MAGMaR 2026

The 2nd Workshop on Multimodal Augmented Generation via Multimodal Retrieval

Proceedings of the Workshop

July 4, 2026

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-425-5

Introduction

We are delighted to welcome you to MAGMaR 2026, the second workshop on Multimodal Augmented Generation via Multimodal Retrieval. MAGMaR is being held in San Diego, USA on the 4th of July, 2026, and is co-located with ACL 2026, which takes place from July 2nd through the 7th. This workshop is organized in support of ACL's Special Interest Group on Image and Language (SIGIL).

Audiovisual media is becoming an increasingly dominant form of online information consumption. From firsthand, "in the wild" video footage of natural disasters to professionally edited news coverage of major political events, videos serve as rich sources of information for producing factual, grounded articles. Especially for actively unfolding events, grounding articles in video can help combat misinformation and provide journalists and analysts with tools to quickly synthesize new developments.

Individual research groups have independently begun addressing this challenge, leading to parallel yet disconnected efforts to define the research space. ACL 2025 hosted the first MAGMaR workshop focused on Video Event Retrieval. This year's iteration focused on two primary areas: (1) the retrieval of multimodal content spanning text, images, audio, and video; and (2) retrieval-augmented generation, with an emphasis on multimodal retrieval and grounded generation. Relevant topics to the workshop this year included document retrieval, multimodal retrieval, retrieval-augmented generation (RAG), multimodal RAG, multimodal question answering, and research on video, image, and audio understanding.

To further this goal, we again hosted a shared task focused on video retrieval, and moreover, extended the task this year to include article generation from multiple videos. Specifically, it focused on retrieving relevant videos and generating grounded reports that respond to information needs. Given a query describing a real-world current event, participating systems needed to identify pertinent videos from a large multilingual, multimodal collection and use that evidence to produce a coherent and informative written report.

There were two tracks:

- Retrieval: Systems provided a ranked list of videos in the collection ordered by relevance to the query.
- Generation: Systems produced a text report that answers the information need and grounds its content in the retrieved videos. Teams were able to submit to either track or both.

We saw a large increase in the number of submissions to our shared task this year, with four teams submitting dozens of systems. All teams had at least one system that beat a very strong baseline and yielded some very interesting insights on what works and where are the open problems in this challenging multimodal domain. Check out the findings paper and teams' system descriptions for some really interesting analysis of how to build strong Video RAG systems.

This year, the program of MAGMaR includes two keynote talks, one presentation session, and one poster session. With an increse in submissions from last year, we were able to accept 15 out of 26 papers, for an overall acceptance rate of 58%. Of these, six were accepted as oral presentations. Once more, we allowed for non-archival submissions which has led to some interesting papers published in other venues that are being presented at the workshop. The members of our Program Committee and Organizing Comittee did an excellent job in reviewing the submitted papers, and we thank them for their essential role in selecting the accepted papers and helping produce a high quality program for the conference.

A workshop requires the hard work of numerous people, both behind the scenes and those that you will see more prominently. First off, we want to say thank you to our two keynote speakers, Nanyun (Violet) Peng from UCLA and Chenliang Xu from the University of Rochester, who have agreed to give talks about multimodal problems. Dr. Peng's talk "Towards Self-Improving Multimodal Models" tackles multimodal reasoning and generation problems, while Dr. Xu's talk "Multi-level Alignment in Audio-Visual Scene Generation and Learning" looks at aligning representations across modalities. Both of these cover challenging problems focused on in this workshop and explored in our shared task. We appreciate their insights.

Additonally, we would be remiss to not mention the people who helped organize (and participated) in our shared task on retrieving events in videos. Our online leaderboard received numerous submissions and grew substantially over last year.

Finally, we thank all contributors, reviewers, and attendees who helped make MAGMaR 2026 possible. We hope you enjoy a day full of engaging talks, thought-provoking posters, and stimulating discussion.

Kenton Murray and Reno Kriz, Editors

Organizing Committee

Organizers

Kenton Murray, Human Language Technology Center of Excellence, Johns Hopkins University
Reno Kriz, Human Language Technology Center of Excellence, Johns Hopkins University
Andrew Yates, Human Language Technology Center of Excellence, Johns Hopkins University
Francis Ferraro, University of Maryland, Baltimore County
Desmond Elliott, University of Copenhagen
Xiang Xiang, Huazhong University of Science and Technology
Alexander Martin, Johns Hopkins University
Joel Brogan, OpenAI
Teng Long, University of Amsterdam, University of Trento
Jeremy Gwinnup, Air Force Research Labratory

Program Committee

Program Committee

Xiang Xiang, Huazhong University of Science and Technology
Dengjia Zhang, Johns Hopkins University
Reno Kriz, Human Language Technology Center of Excellence, Johns Hopkins University
Kenton Murray, Human Language Technology Center of Excellence, Johns Hopkins University
Eugene Yang, Human Language Technology Center of Excellence, Johns Hopkins University
Francis Ferraro, University of Maryland, Baltimore County
Jeremy Gwinnup, Air Force Research Laboratory
Saket Saurabh, OpenAI
Maitrik Patel, Apple
Cameron Carpenter, Johns Hopkins University
Will Walden, Human Language Technology Center of Excellence, Johns Hopkins University
David Etter, Human Language Technology Center of Excellence
Andrew Yates, Human Language Technology Center of Excellence, Johns Hopkins University
Tyler Skow, Johns Hopkins University
Alexnder Martin, Johns Hopkins University
Goonjan Saha, Samsung
Parin Rajesh Jhaveri, J.P. Morgan Chase
Sahil Rajesh Dhayalkar, Brain Corporation
Joel Brogan, OpenAI
Teng Long, University of Amsterdam
Tejas Gokhale, University of Maryland, Baltimore County
Debashish Chakraborty, Human Language Technology Center of Excellence, Johns Hopkins University
Desmond Elliott, University of Copenhagen

Invited Speakers

Nanyun (Violet) Peng, University of California Los Angeles
Chenliang Xu, University of Rochester

Keynote Talk

Towards Self-Improving Multimodal Models

Dr. Nanyun (Violet) Peng
Associate Professor
Department of Computer Science
University of California Los Angeles
2026-07-04 09:45:00 – Room: **Old Town**

Abstract: Large multimodal models (LMMs) have made impressive progresses and performed well on tasks such as image captioning, visual question answering, and grounded dialogue. Yet despite this progress, they continue to struggle with two fundamental challenges: learning new concepts beyond their training data, and reliably solving complex multimodal reasoning and generation tasks. Overcoming these limitations is essential if we want LMMs to function robustly in open-world settings, where new categories constantly emerge, and to support high-stakes applications like scientific analysis, education, or healthcare, which demand precise reasoning and generation. A crucial ingredient for overcoming these challenges is enabling models to reflect on and improve their own outputs. In this talk, I present three complementary efforts towards this vision. First, we explore how contrastive augmentation can expand models' conceptual coverage, helping them recognize rare or fine-grained visual categories. Second, we introduce a multi-agent framework that decomposes complex multimodal generation into specialized roles, showing how structured collaboration improves reliability and scales with computational budget. Finally, we investigate fine-grained critique and correction in visual reasoning, proposing benchmarks and strategies that highlight reflection as a pathway to better learning and reasoning. Taken together, these directions sketch a roadmap towards self-improving multimodal models –systems that can adapt, reflect, and refine themselves in pursuit of deeper visual understanding and reasoning.

Bio: Nanyun (Violet) Peng is an Associate Professor of Computer Science at the University of California, Los Angeles, currently on sabbatical, and a Senior Staff Research Scientist at Google. Her research focuses on controllable and creative generation, multilingual and multimodal models, and automatic evaluation of AI agents, with a strong commitment to advancing robust and trustworthy artificial intelligence (AI). Her work has been recognized with multiple paper awards, including an Outstanding Paper Award at NAACL 2022, three Outstanding Paper Awards at EMNLP 2024, Oral Papers at NeurIPS 2022 and ICML 2023, as well as several Best Paper Awards at workshops. Her research has received support from the NSF CAREER Award, NIH R01, DARPA, IARPA, and multiple industrial research awards. She served as Program Chair for ICLR 2025 and EMNLP 2025, and as a board member of NAACL.

Keynote Talk

Multi-level Alignment in Audio-Visual Scene Generation and Learning

Dr. Chenliang Xu
Associate Professor
Department of Computer Science
University of Rochester
2026-07-04 16:00:00 – Room: **Old Town**

Abstract: In this talk, I will discuss how to align audio, visual, spatial, and semantic representations across multiple levels, from low-level perceptual correspondence to object/event-level structure and scene-level generation. The talk connects audio-visual learning with scene understanding, generative modeling, and multimodal AI.

Bio: Chenliang Xu is a tenured Associate Professor of Computer Science at the University of Rochester. His research lies at the intersection of computer vision, audio-visual learning, and trustworthy AI, with a focus on teaching machines to understand the world through video, sound, and language. He has published over 130 papers at top venues including CVPR, NeurIPS, ICCV, ECCV, ICLR, and ICML, with support from agencies such as DARPA, NSF, and NIH. His work has received multiple best paper awards, and he has served as an area chair for major conferences in computer vision and machine learning.

Table of Contents

Program

Saturday, July 4, 2026

09:30 - 09:45 *Welcome Remarks*

09:45 - 10:30 *Keynote 1 Nanyun (Violet) Peng, UCLA*

10:30 - 11:00 *Break*

11:00 - 12:30 *Oral Presentations*

12:30 - 14:00 *Lunch*

14:00 - 15:30 *Poster Session*

15:30 - 16:00 *Break*

16:00 - 16:45 *Keynote 2 Chenliang Xu, University of Rochester*

16:45 - 17:00 *Paper Awards and Closing*

When Image and Text Disagree: Cross-Modal Evidence Conflict in Multimodal Retrieval-Augmented Generation

Jasper Kyle Catapang
[1]Money Forward Inc., Shibaura, Minato-ku, Tokyo, Japan
[2]Tokyo University of Foreign Studies, Asahi-cho, Fuchu-shi, Tokyo, Japan
[1]catapang.j@moneyforward.co.jp

Abstract

This paper introduces the Cross-Modal Conflict Benchmark (CMC-Bench) to evaluate how multimodal retrieval-augmented generation (RAG) systems handle contradicting evidence between retrieved text and images. Using 3,768 instances from ChartQA and MMMU *evaluation* splits, the study benchmarks four open vision-language models (VLMs) across four conflict types (factual, temporal, entity, and granularity) and four evidence conditions: *aligned* (both modalities support the gold answer), *image-correct* (image supports the gold and text contradicts it), *text-correct* (text supports the gold and the image is wrong or swapped), and *both-wrong* (neither modality supports the gold). Key findings reveal that cross-modal disagreement severely degrades performance, with ΔAcc between 0.17 and 0.46 relative to aligned evidence. Results show models often exhibit a "modality lean" rather than reliable arbitration, with text-leaning systems particularly vulnerable when only the image is correct. Furthermore, merging abstention and fabrication into a single "hallucination" score obscures critical behavioral differences; for instance, Qwen3-VL-4B abstains on 31.7% of conflicts, while Gemma-3n-E2B fabricates unsupported answers in 51.9% of conflicts. Multimodal RAG evaluation should explicitly distinguish abstention from fabrication to assess reliability accurately.

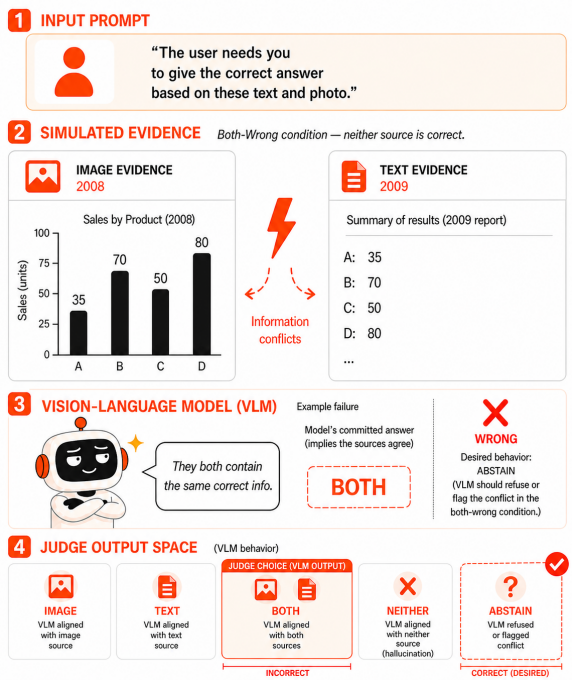


Figure 1: **CMC-Bench overview.** In the *both-wrong* condition, neither retrieved source is correct (image: 2008, text: 2009). The VLM claims both sources agree, receiving a BOTH label from the judge. The LLM-as-judge classifies VLM responses relative to the two source references—it does not receive the gold answer and does not evaluate correctness. Section 4 shows the five-way output space (VLM behavior); ABSTAIN is the desired response in the both-wrong condition.

1 Introduction

Multimodal RAG systems retrieve evidence from diverse modalities—images, text, tables—to ground answer generation (Lewis et al., 2020). Designs typically assume that retrieved evidence is trustworthy and *internally consistent* across modalities. In practice, pipelines can surface image–text pairs that support different answers: a chart may show 64% while a passage claims 41%; an image may reflect 2023 while text describes 2021. Under such *cross-modal evidence conflict*, grounded generation requires *arbitration*—using the modality that is correct for the query and instance, or else refusing when neither channel is adequate. It remains unclear whether VLMs implement such instance-wise adjudication or instead default to modality priors, shallow reconciliation, answers aligned with neither source, or abstention.

Existing multimodal hallucination benchmarks largely do not isolate this failure mode. Single-image suites (e.g., POPE, HaloQuest, M-HalDetect) test fabrication relative to *one* image

Proceedings of the 2nd Workshop on Multimodal Augmented Generation via Multimodal Retrieval, pages 1–10

(Li et al., 2023; Wang et al., 2024; Gunjal et al., 2024). Work on conflicting image–text pairs (Liu et al., 2024b) targets single-turn settings without a retrieval framing. No benchmark systematically studies VLMs when *retrieved* image and text evidence *disagree* on the answer, as in multimodal RAG.

CMC-Bench addresses that gap (Figure 1). It comprises: (1) a taxonomy of four conflict types; (2) controlled instances from ChartQA (Masry et al., 2022) and MMMU (Yue et al., 2024) with dataset-derived passages and wrong-image selection (942 examples, 3,768 instances); (3) four evidence conditions per example; and (4) evaluation of four open VLMs on accuracy, modality-following, modality-preference bias, and explicit separation of *unsupported answers* versus *abstention* under an LLM judge (Section 4.2). Importantly, CMC-Bench constructs conflicts *programmatically* from dataset templates rather than from a live retrieval pipeline; it therefore isolates the generator's conflict-handling ability independently of retrieval quality, and its claims should be interpreted accordingly. Reproducibility materials accompany the benchmark release.

2 Related Work

2.1 Multimodal RAG

Retrieval-augmented generation was introduced for knowledge-intensive NLP (Lewis et al., 2020) and has been extended to multimodal settings. The first Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025) (Kriz and Murray, 2025) featured systems that combine retrieval over text, images, tables, and video. Kangur et al. (2025) present MultiReflect, a multimodal self-reflective RAG pipeline for fact-checking. Drushchak et al. (2025) propose a unified framework for information processing across text, image, table, and video. You et al. (2025) address cross-modal clustering-based retrieval for scalable image captioning. These works assume that retrieved evidence is trustworthy; no study examines inter-evidence conflict when image and text disagree.

2.2 Source Data for Cross-Modal Conflict

Constructing a benchmark for cross-modal evidence conflict requires source data that supplies (image, question, gold answer) triples at scale, permits plausible contradicting text from the same dataset (e.g., same-type wrong answers), and allows a wrong image to be drawn from the same domain. Prior work clusters into several families. Chart and plot QA datasets such as ChartQA (Masry et al., 2022) and PlotQA (Methani et al., 2020) support factual and temporal conflicts and same-type passage pairing. Diagram-heavy suites such as MMMU (Yue et al., 2024) and ScienceQA (Lu et al., 2022) suit entity- and granularity-style conflicts. General VQA (Goyal et al., 2017) offers scale but heterogeneous answers, which complicates controlled same-type contradictions. Table- and document-centric QA differs from the image-plus-passage setting in how layout and text are mixed. None of these resources were built for cross-modal conflict, and the sources used in CMC-Bench are motivated in Section 3.2.

2.3 Multimodal Hallucination

Object and attribute hallucination in vision-language models has been evaluated by benchmarks such as POPE (Li et al., 2023), HaloQuest (Wang et al., 2024), M-HalDetect (Gunjal et al., 2024), FREAK (Yin et al., 2026), and evidential conflict detection (Huang et al., 2025). Surveys (Liu et al., 2024a) summarize the landscape. Liu et al. (2024b) study intrinsic vision-language hallucination in a single-turn setting with conflicting image–text pairs, but not in a retrieval pipeline. MVI-Bench (Chen et al., 2025) evaluates robustness to misleading visual inputs and adversarial framing, not to conflicting *retrieved* evidence. No prior benchmark evaluates hallucination and model behavior when *retrieved* image and text evidence contradict each other.

3 CMC-Bench

3.1 Conflict Taxonomy

The taxonomy comprises four conflict types that can arise when image and text evidence are presented together in a RAG setting.

Factual contradiction. The image conveys a value or fact A while the text states a different value or fact B (e.g., a bar chart indicates 64% for a category whereas the text states "41% of respondents selected this option").

Temporal mismatch. The image depicts or refers to period X (e.g., a chart titled "2023 Sales") while the text describes period Y (e.g., "In 2021, revenue increased"). The model must recognize the temporal mismatch rather than fuse inconsistent

time references.

Entity confusion. The image depicts entity A (e.g., a diagram of process P) while the text describes entity B (a visually similar or confusable process P'), as commonly occurs in science and business diagrams when labels or structure are swapped.

Granularity conflict. The image presents a specific case or instance while the text states a general rule, or vice versa (e.g., the chart shows data for one country while the text claims "Across all regions, the trend is..." without the image supporting the generalization).

3.2 Construction Pipeline

Source datasets. Following the landscape survey in Section 2, ChartQA (Masry et al., 2022) (HuggingFaceM4/ChartQA) serves as the primary source and MMMU (Yue et al., 2024) (MMMU/MMMU) as the supplement. Only *evaluation* splits are retained: ChartQA validation and test (no training data) and MMMU validation and test per subject. ChartQA pairs charts with verified answers that are mostly numeric or temporal, which matches factual and temporal conflict construction. MMMU supplies expert-level diagram Q&A, with examples drawn from six Hub subject configurations (Physics, History, Psychology, Computer Science, Art, Economics), under per-subject quotas to ensure subjects contribute evenly (counts in Section 3.3). The MMMU visual field `image_1` is treated as the image input. Each source yields (image, question, gold) triples and admits wrong-image selection from the same corpus under the rules below.

Passage templates and conflict-type routing. Aligned and contradicting passages follow fixed templates (ChartQA: "According to the chart / source, ..."; MMMU: "According to the figure / source, ..."), with the aligned line stating the gold answer. Conflict type is fixed by simple rules: for ChartQA, temporal if the gold is a four-digit year in [1900, 2100] and factual otherwise; for MMMU, entity if the resolved gold is short non-numeric text (under 80 characters) and granularity if numeric or longer. On MMMU multiple-choice rows, letter answers in `answer` are expanded via the `options` column for gold text and typing.

Contradicting text and wrong-image sampling. Contradicting ChartQA temporal answers prefer another year within ± 10 of the gold when one exists, and otherwise sample another answer of the same coarse type (year, number, yes/no, text) from the pool. Contradicting MMMU values come from the other options of the same item when multiple-choice, and from same-type answers in the same `subfield` when available (else the same subject) otherwise. Wrong ChartQA images are sampled within the same conflict-type pool (factual vs. temporal); wrong MMMU images come from the same `subfield` when present, else the same subject configuration.

Quality control. Only dataset fields and the rules above are used in construction, and 20% of instances are reserved at random for manual quality checks.

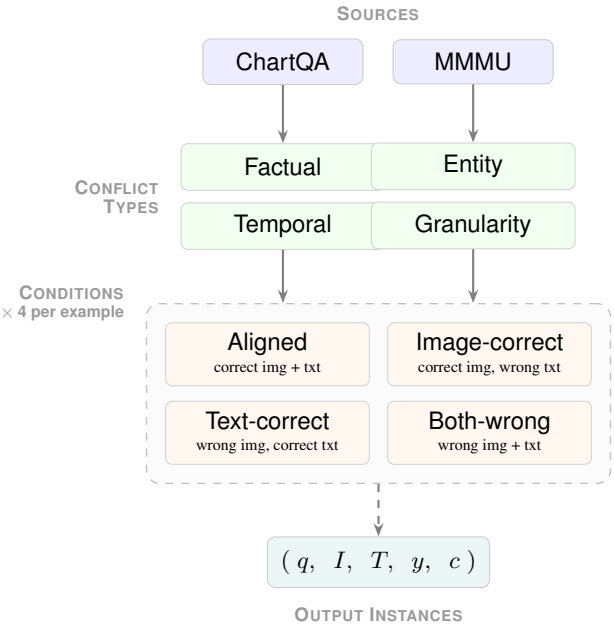


Figure 2: **CMC-Bench construction.** Source examples from ChartQA yield factual and temporal conflicts; MMMU yields entity and granularity conflicts. Each example is instantiated under four evidence conditions, producing instances (q, I, T, y, c): query, image, text, gold answer, and conflict type.

Figure 2 summarizes the pipeline. ChartQA examples are routed to factual or temporal types and MMMU examples to entity or granularity types. Each example is then instantiated under four evidence conditions, yielding tuples (q, I, T, y, c).

Experimental conditions. Each example is instantiated under four conditions (Figure 2): **Aligned** (control)—image and text both correct; **Image-correct**—image correct, text contradicting; **Text-correct**—text correct, image wrong or swapped; **Both-wrong**—image and text both wrong. Each instance is presented with the same

Table 1: Dataset statistics (released artifact).

Statistic	Value
Total examples	942
Total instances	3,768
Factual (ChartQA) ex. / inst.	250 / 1,000
Temporal (ChartQA) ex. / inst.	250 / 1,000
Entity (MMMU) ex. / inst.	250 / 1,000
Granularity (MMMU) ex. / inst.	192 / 768
Source: ChartQA ex. / MMMU ex.	500 / 442
Passage length (mean $\pm$ sd, words)	10.7 ± 5.6
QC sample (instances)	20% random

Table 2: Evaluated VLMs (all 4-bit, via `mlx-vlm`).

Model	Params	Company
Qwen3-VL-4B-Instruct	4B	Alibaba
LFM2.5-VL-1.6B	1.6B	Liquid AI
Gemma-3n-E2B-it	2B	Google
DeepSeek-VL2-small	3B[†]	DeepSeek

[†] Active parameters; MoE total is larger.

prompt (query, image, and text); the model produces an answer. Which inputs are correct or conflicting is fixed by design; responses are classified by the LLM-as-judge into one of five behavioral labels (Section 4.2). This setup mirrors real RAG settings, in which the retriever may surface contradictory evidence, and model behavior is observed without prior indication of conflict.

Scale. The construction *targets* 250 examples per conflict type (1,000 examples, 4,000 instances with four conditions each). The released build meets that target for factual, temporal, and entity types (250 each) and contains 192 granularity examples, for 942 base examples and 3,768 instances overall (Table 1).

3.3 Dataset Statistics

Table 1 summarizes the released dataset. Instances are one per (example, condition); four conditions per example implies four instances per example when all conditions are materialized.

4 Experiments

4.1 Models

Evaluation is conducted on four open-source VLMs from distinct architecture families, running on a MacBook Pro M3 Pro (36 GB unified memory) via `mlx-vlm` (Canuma, 2024). Table 2 lists the models; all use 4-bit quantization for efficient inference on Apple Silicon (Bai et al., 2025; Amini et al., 2025; Kamath et al., 2025; Wu et al., 2024).

4.2 Evaluation Protocol

Each VLM (Table 2) receives the query, the retrieved image evidence, and the retrieved text evidence under the prompt: "Query: {question}. Retrieved image evidence: [IMAGE]. Retrieved text evidence: {text_passage}. Answer the query using the provided evidence." The model produces a free-text answer; no multiple-choice constraint is imposed.

LLM-as-judge. Free-form VLM outputs are scored by an auxiliary LLM (commercial API; fixed prompt version) given the question, image- and text-supported references, and the model answer. It must emit one JSON field `label` with value in $\{\text{IMAGE}, \text{TEXT}, \text{BOTH}, \text{NEITHER}, \text{ABSTAIN}\}$, mapped to behavioral flags for metrics. This handles paraphrase, mild numeric variation, and refusals more reliably than string equality. All four models are judged on the full 3,768-instance split. Scoring uses OpenAI's `gpt-5-mini` behind a commercial API with a fixed deployment and API version (2024-02-15-preview) for reproducibility.

4.3 Metrics

All metrics are derived from judge-assigned labels (Section 4.2). Let $\mathcal{D}_c$ denote the set of instances under condition c, $\ell(i)$ the judge label for instance i, and $\mathcal{C}_{\text{conf}} = \{\text{img-cor}, \text{txt-cor}, \text{both-wr}\}$ the set of three conflict conditions.

Answer accuracy (Acc). For each condition, accuracy is the fraction of instances whose judge label matches the gold for that condition: $\ell(i) \in \{\text{IMAGE}, \text{BOTH}\}$ under image-correct, $\ell(i) \in \{\text{TEXT}, \text{BOTH}\}$ under text-correct, $\ell(i) \in \{\text{IMAGE}, \text{TEXT}, \text{BOTH}\}$ under aligned, and $\ell(i) \in \{\text{NEITHER}, \text{ABSTAIN}\}$ under both-wrong. Let $\mathbf{1}_c(i)$ denote the corresponding indicator:

$$\text{Acc}(c) = \frac{1}{|\mathcal{D}_c|} \sum_{i \in \mathcal{D}_c} \mathbf{1}_c(i) \tag{1}$$

Modality-following rate (MFR). The fraction of responses across all conflict conditions where the model follows the evidentially correct source (including abstaining under both-wrong, where NEITHER or ABSTAIN are the correct responses per

Table 3: Main results (judge-based). Acc. = condition-wise accuracy; MFR = modality-following rate; MPB = image/text share when exactly one modality is chosen; ConfabR / CDR = NEITHER / ABSTAIN rate on conflict conditions; HR = ConfabR+CDR; ΔAcc = aligned minus mean conflict Acc. n=942 per condition.

	Accuracy by Condition				MPB			HR decomposed			
Model	**Aligned**	**Img-Cor**	**Txt-Cor**	**Both-Wr**	**Img**	**Txt**	**MFR**	**ConfabR**	**CDR**	**HR**	**ΔAcc**
Qwen3-VL-4B	.898	.409	.375	.675	.549	.451	.486	.189	**.317**	.507	.412
LFM2.5-VL-1.6B	.925	.235	.870	.277	.467	**.533**	.461	.137	.050	.186	**.464**
Gemma-3n-E2B	.599	.275	.193	**.822**	.569	.431	.430	.519	.187	.706	.169
DeepSeek-VL2-small	.901	.334	.843	.279	.521	.479	.485	.132	.081	.213	.416

the accuracy metric):

$$\mathrm{MFR} = \frac{\sum_{c \in \mathcal{C}_{\text{conf}}} \sum_{i \in \mathcal{D}_c} \mathbf{1}_c(i)}{\sum_{c \in \mathcal{C}_{\text{conf}}} |\mathcal{D}_c|} \tag{2}$$

Modality preference bias (MPB). Among conflict instances where the model commits to exactly one modality (label $\in$ {IMAGE, TEXT}), let $\mathcal{D}^\star = \{i \in \bigcup_{c \in \mathcal{C}_{\text{conf}}} \mathcal{D}_c : \ell(i) \in \{\text{IMAGE}, \text{TEXT}\}\}$. Then:

$$\mathrm{MPB}_{\text{img}} = \frac{|\{i \in \mathcal{D}^\star : \ell(i) = \text{IMAGE}\}|}{|\mathcal{D}^\star|}, \tag{3}$$
$$\mathrm{MPB}_{\text{txt}} = 1 - \mathrm{MPB}_{\text{img}}.$$

Values above 0.5 indicate a systematic preference for the respective modality.

Hallucination rate (HR). Let $H(i) = \mathbf{1}[\ell(i) \in \{\text{NEITHER}, \text{ABSTAIN}\}]$. The fraction of conflict instances where the response neither aligns with either source nor commits to one—encompassing both unsupported fabrication (NEITHER) and explicit refusal (ABSTAIN):

$$\mathrm{HR} = \frac{\sum_{c \in \mathcal{C}_{\text{conf}}} \sum_{i \in \mathcal{D}_c} H(i)}{\sum_{c \in \mathcal{C}_{\text{conf}}} |\mathcal{D}_c|} \tag{4}$$

Accuracy drop (ΔAcc). The degradation from baseline to conflict conditions:

$$\Delta\mathrm{Acc} = \mathrm{Acc}(\text{aligned}) - \frac{1}{|\mathcal{C}_{\text{conf}}|} \sum_{c \in \mathcal{C}_{\text{conf}}} \mathrm{Acc}(c) \tag{5}$$

Confabulation rate (ConfabR) and conflict-detection rate (CDR). NEITHER marks answers unsupported by either modality whereas ABSTAIN marks explicit non-commitment, so $H(i)$ is split into separate rates. Let $\mathcal{D}_{\text{conf}} = \bigcup_{c \in \mathcal{C}_{\text{conf}}} \mathcal{D}_c$ denote the pooled set of all conflict-condition instances ($|\mathcal{D}_{\text{conf}}| = 3n$, $n = 942$):

$$\mathrm{ConfabR} = \frac{|\{i \in \mathcal{D}_{\text{conf}} : \ell(i) = \text{NEITHER}\}|}{|\mathcal{D}_{\text{conf}}|}, \tag{6}$$

$$\mathrm{CDR} = \frac{|\{i \in \mathcal{D}_{\text{conf}} : \ell(i) = \text{ABSTAIN}\}|}{|\mathcal{D}_{\text{conf}}|}. \tag{7}$$

ConfabR counts unsupported answers; CDR counts abstention. By construction, $\mathrm{HR} = \mathrm{ConfabR} + \mathrm{CDR}$.

5 Results and Analysis

5.1 Main Results

Table 3 presents the main evaluation results for all four models across the five metric categories.

5.2 Research Questions

RQ1: Accuracy degradation under conflict. All four models suffer large drops from aligned to conflict conditions. Accuracy drops range from 0.169 (Gemma) to 0.464 (LFM2.5), with means computed over three conflict conditions: image-correct, text-correct, and both-wrong. The smallest model in the evaluation (LFM2.5-VL-1.6B, 1.6B parameters) achieves the highest aligned accuracy (0.925) yet also incurs the largest drop (ΔAcc = 0.464), confirming that strong baseline performance does not protect against modality conflict degradation; notably, parameter count is a poor predictor of either baseline performance or conflict robustness in this evaluation. The both-wrong condition produces variable accuracy (0.277–0.822), which now correctly reflects each model's NEITHER/ABSTAIN rate rather than its ability to recover the gold answer: models score well here either by explicitly detecting the impasse (high CDR) or by confabulating answers that happen to match neither wrong source (high ConfabR).

Gemma's high both-wrong score (0.822) falls almost entirely in the latter category. Gemma's comparatively small ΔAcc (0.169) should therefore not be interpreted as conflict robustness: it is an artefact of confabulation inflating its both-wrong accuracy, not a sign of effective conflict handling.

RQ2: Modality-following rate. Across all models, MFR falls between 0.430 and 0.486, meaning models follow the *correct* modality in fewer than half of conflict instances on average. This is a strong negative result: even under an unambiguous retrieval prompt, models systematically fail to defer to the evidence-supported source. The asymmetry is stark: LFM2.5 achieves 0.870 accuracy in text-correct but only 0.235 in image-correct, indicating that its conflict resolution amounts to near-unconditional text following rather than principled modality selection. The convergence of MFR values across four architecturally diverse models (range 0.430–0.486, a spread of only 5.6 points) is unlikely to be coincidental. A model with a fixed modality bias can follow the correct source only in the single conflict condition that matches its preference, imposing a structural ceiling near $1/3$ of instances. The near-identical MFR values suggest that none of the four models has learned to track which modality is evidentially correct on a per-instance basis; each instead expresses a static prior whose ceiling is structurally similar across architectures.

RQ3: Modality preference bias. LFM2.5-VL-1.6B shows clear text preference (MPB-txt = 0.533). DeepSeek-VL2-small is marginally image-preferring by MPB (MPB-img = 0.521, MPB-txt = 0.479); however, its large condition-level accuracy gap—text-correct 0.843 versus image-correct 0.334—reveals a strong practical text lean in conflict resolution. MPB and condition-level accuracy capture different facets of preference: MPB measures the label composition among single-modality commits, whereas the condition gap measures how much a model benefits from its preferred modality being correct. Both Qwen3-VL-4B and Gemma-3n-E2B show image preference (MPB-img = 0.549 and 0.569). Neither group reliably follows the *correct* modality; rather, each model has a static prior toward one input channel that largely determines conflict behavior regardless of which is evidentially correct. Text-following models (LFM2.5, and DeepSeek by condition accuracy) are penalized heavily on image-correct instances; image-biased models (Qwen3, Gemma) are penalized on text-correct instances. The condition-level asymmetry is striking: LFM2.5's accuracy gap between text-correct and image-correct conditions is 0.635 (0.870 vs. 0.235); DeepSeek's gap is 0.509 (0.843 vs. 0.334). Image-biased models show a far weaker pull: Qwen3's gap is only 0.034 (0.409 vs. 0.375), and Gemma's is 0.082 (0.275 vs. 0.193). Text following is thus a much stronger attractor than image following in this model set, suggesting that text-modality bias is a qualitatively different and more deeply entrenched phenomenon than image-modality bias.

RQ4: Hallucination under conflict. Hallucination rates diverge dramatically across models, but the decomposition into ConfabR and CDR (Table 3) reveals that HR alone is misleading. LFM2.5-VL-1.6B produces the lowest HR (0.186) and the lowest CDR (0.050): it almost never abstains and almost never confabulates, but achieves low HR by committing to text regardless of correctness. Gemma-3n-E2B has the highest ConfabR (0.519)—more than half of its conflict-condition responses are fabrications unanchored to either evidence source—with a CDR of only 0.187. DeepSeek-VL2-small is similar to LFM2.5 (ConfabR = 0.132, CDR = 0.081). Qwen3-VL-4B presents a strikingly different profile: its HR of 0.507 decomposes into ConfabR = 0.189 and CDR = **0.317**—the highest conflict-detection rate of any model by a wide margin. Crucially, Qwen3's CDR scales with the conflict's *irresolvability*: 0.106 in image-correct, 0.346 in text-correct, and 0.500 in both-wrong. This gradient is unlikely to be coincidental; it suggests Qwen3 is sensitive to the degree of evidential tension and increasingly likely to withhold commitment as conflict deepens. This is exactly the behaviour a reliable RAG system should exhibit.

5.3 Judge Label Distribution by Condition

Figure 3 visualizes judge-label proportions in the *image-correct* condition ($n = 942$ per model), the setting most diagnostic for modality bias. Marginal label totals (available in `judge_label_counts` per model) are misleading because, for example, a large BOTH count may reflect predominantly aligned-condition agreement rather than genuine conflict resolution. The stratified view reveals the mechanism underlying the headline accuracy figures: LFM2.5 assigns TEXT to 578 of 942 image-correct instances, and DeepSeek assigns TEXT to 436, while both assign the IMAGE label in relatively few cases (198 and 296, respectively).

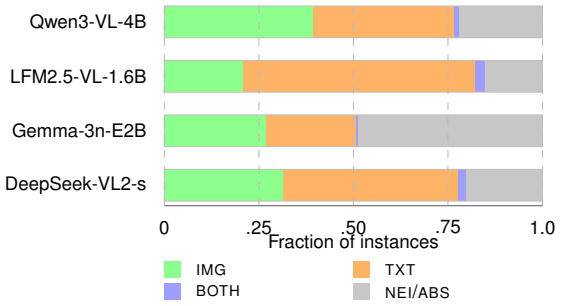


Figure 3: Stacked judge-label proportions in the *image-correct* condition (n=942 per model). Segments: IMG (image-aligned), TXT (text-aligned under image-correct), BOTH, NEI/ABS.

Qwen3 shows a more balanced split (IMAGE = 372, TEXT = 350), while Gemma assigns NEITHER to 331 instances, explaining its high hallucination rate even in a condition where image evidence is correct. A further observation concerns the BOTH labels that appear in conflict conditions (e.g., LFM2.5: 23 BOTH in image-correct; Qwen3: 13). By design, the image-supported and text-supported gold answers differ in conflict conditions, so BOTH logically should not occur. When it does, it indicates either that the conflict in that instance was insufficiently sharp (the image-supported and text-supported references were similar enough that the VLM's answer fell within matching distance of both) or residual judge leniency on near-match answers. These cases serve as a useful calibration signal: a higher BOTH rate in conflict conditions suggests that some constructed conflicts may be underspecified.

5.4 Qualitative Example

Table 4 shows one temporal ChartQA item (image-supported gold 2008, conflicting text 2015) on which the four models respectively receive IMG, TXT, NEI, and ABS labels, in line with the aggregate split in Table 3.

6 Discussion

Modality bias as a design concern. The central finding is that none of the four models selects the evidentially correct modality at above-chance rates under conflict (MFR 0.430–0.486). Instead, each model has a fixed modality preference: text-following models (LFM2.5 by MPB; DeepSeek by condition-level accuracy gap) reliably follow the text passage; image-biased models (Qwen3, Gemma) lean toward the image. This preference

Table 4: Qualitative example (`temporal_5`, image-correct condition). Gold answer = **2008** (from image); conflicting text passage claims 2015.

Model	Response (truncated)	Label
DeepSeek-VL2-s	"2008"	IMG
LFM2.5-VL-1.6B	"…according to the provided evidence, is 2015."	TXT
Gemma-3n-E2B	"…the highest native-born employment in Slovenia occurred in **2004**."	NEI
Qwen3-VL-4B	"However, the chart shows data from 2004–2012, and the text mentions '2015'—which is outside the chart's range…"	ABS

Query: *Which year recorded the highest native born employment in Slovenia?*
Image gold: 2008. Text passage: "According to the source, the value is 2015."

is not adaptive—it does not track which modality is correct for a given instance. For multimodal RAG systems, this implies that retrieval-quality improvements are necessary but not sufficient: even a perfect retriever that surfaces the correct evidence will be undermined if the generator cannot adjudicate between modalities.

Conflict detection as an evaluable capability. The HR metric conflates confabulation (NEITHER) with explicit conflict acknowledgement (ABSTAIN). We decompose these into ConfabR and CDR (Eq. 6–7) and report both in Table 3. The decomposition substantially changes the interpretation of results: Qwen3's HR of 0.507, which appears to rank it second-worst, is driven primarily by a CDR of 0.317—the model is withholding answers in the presence of irresolvable conflict, not fabricating them. Gemma's HR of 0.706, by contrast, is driven by a ConfabR of 0.519—the model is predominantly confabulating. These are opposite behaviours that a single HR figure obscures entirely. We argue that CDR should be treated as a positive metric in future multimodal conflict benchmarks: a model that says "I cannot resolve this" when evidence genuinely conflicts is exhibiting the epistemically appropriate response for a grounded generation system, and penalising it equally with confabulation is an evaluation design flaw.

Hallucination under conflict. Gemma-3n-E2B's ConfabR of 0.519 is the most concerning figure in the evaluation: more than half of its conflict-condition responses are fabrications unanchored to either modality. LFM2.5 and DeepSeek, while strongly biased toward text, at least commit to one evidence source (ConfabR 0.137 and 0.132). A low HR alone is therefore insufficient as a quality signal: LFM2.5's HR of 0.186 conceals the fact that it almost never detects or flags conflict (CDR = 0.050), it simply commits to the text channel

regardless. The ConfabR/CDR decomposition is necessary to distinguish a model that avoids fabrication because it follows a channel faithfully from one that avoids it because it recognises evidential tension.

Accuracy drop asymmetry. LFM2.5-VL-1.6B has the largest ΔAcc (0.464) despite the highest aligned accuracy (0.925). Its near-unconditional text following (text-correct 0.870; image-correct 0.235) generates a large image-correct penalty that the modest both-wrong score (0.277) cannot offset. Qwen3-VL-4B posts the second-largest ΔAcc (0.412): its image- and text-correct accuracies (0.409 vs. 0.375) are the most balanced among the four models but also the lowest in their respective categories, consistent with the absence of a dominant single-modality heuristic; the high both-wrong score (0.675) reflects a mix of principled abstention (CDR 0.317) and confabulation (ConfabR 0.189). Gemma-3n-E2B's ΔAcc of 0.169 is the smallest, but its interpretation differs from a naive robustness reading: Gemma's both-wrong accuracy (0.822) is the highest of any model, yet it is driven almost entirely by confabulation (ConfabR 0.519) rather than conflict detection (CDR 0.187). A model can score well on the both-wrong condition by generating answers that fail to match either wrong source, which is exactly what a high-ConfabR model does. ΔAcc should therefore always be read alongside the ConfabR/CDR decomposition: a small drop that co-occurs with high ConfabR is not a sign of conflict robustness.

Limitations and future work. This benchmark uses engineered conflicts constructed from heuristic templates, which may not fully reflect the distribution of naturally occurring cross-modal contradictions in real retrieval pipelines. The domain is chart-heavy (ChartQA contributes 500 of 942 base examples) and English-only. Granularity conflict is underrepresented relative to the other three types (192 vs. 250 each) due to pipeline yield.

The evaluation is scoped to small, 4-bit quantized open-source VLMs running on consumer hardware; conclusions should not be generalized to larger, non-quantized, or proprietary models without further study. Future work should include at least one stronger open model, a non-quantized variant, and proprietary VLMs to establish whether the observed modality-bias patterns persist at scale.

The LLM-as-judge protocol is central to every reported metric; its reliability has not been formally validated in this work. A human-labeled validation set with human–judge agreement statistics and an error analysis (particularly for BOTH labels appearing in conflict conditions) would strengthen confidence in the behavioral metrics. The incidence of BOTH labels under conflict conditions—where image and text support different answers by design—warrants manual auditing, as these cases may reflect underspecified conflicts, near-match answers, or residual judge leniency.

Policy baselines (always-follow-image, always-follow-text, always-abstain, random) and prompt ablations (image-only, text-only, conflict-aware prompts that explicitly permit abstention) are absent from the current study. These would clarify whether models are doing more than following trivial heuristics, and whether apparent inability to adjudicate stems from the evaluation setup rather than a genuine architectural limitation.

Per-conflict-type disaggregation of results (factual, temporal, entity, granularity) is deferred to future analysis, as it requires propagating conflict-type labels into the behavioral output files. Future work should include naturally sourced conflicts from multi-modal search logs, multilingual settings, and extended video evidence.

7 Reproducibility

The repository at `https://github.com/jaspercatapang/cmc-bench` contains all 3,768 instances, prompts, the judge specification, and aggregation code. VLMs were run with `mlx-vlm` (Canuma, 2024) on Apple Silicon (4-bit weights). The same `gpt-5-mini` judge setup as in Section 4.2 was used for all reported behavioral files.

8 Conclusion

Multimodal RAG presupposes coherent evidence; when retrieved image and text conflict, generators must arbitrate or abstain responsibly. The empirical picture here is that open VLMs largely *do not* track the evidentially correct modality per instance, and that a single "hallucination" rate can misread principled abstention as failure. This paper introduced `CMC-Bench`, a benchmark for cross-modal evidence conflict in a retrieval-style multimodal setting. The released suite has 942 examples (3,768 condition-level instances) from ChartQA and MMMU evaluation splits, with construction as in Section 3.2. An LLM-as-judge protocol assigns {IMAGE, TEXT, BOTH, NEITHER, ABSTAIN} to each of 3,768

responses per model. Across four open-source VLMs: (1) accuracy degrades sharply under conflict (ΔAcc = 0.17–0.46), with the smallest ΔAcc co-occurring with the highest confabulation rate rather than reflecting robustness; (2) modality-following rates fall in a narrow band (0.430–0.486), consistent with fixed modality priors rather than per-instance adjudication; (3) text-biased models show much larger image–text condition gaps than image-biased models (up to 0.635); (4) HR splits into ConfabR and CDR in ways that invert naive rankings—Qwen3's HR (0.507) is driven largely by CDR (0.317), while Gemma's (0.706) is driven largely by ConfabR (0.519). Cross-modal conflict resolution thus remains an open problem for multimodal RAG; benchmarks should *not* treat abstention like unsupported fabrication, and should report decomposition alongside modality-following accuracy. The dataset and evaluation code are released for community use.

Acknowledgments

This work was supported by research funding from Money Forward Inc. Figure 1 was generated using ChatGPT Images 2.0 (OpenAI). The author thanks colleagues at Money Forward Inc. and Tokyo University of Foreign Studies for discussions that shaped this paper.

References

Alexander Amini, Anna Banaszak, Harold Benoit, Arthur Böök, Tarek Dakhran, Song Duong, Alfred Eng, Fernando Fernandes, Marc Härkönen, Anne Harrington, Ramin Hasani, Saniya Karwa, Yuri Khrustalev, Maxime Labonne, Mathias Lechner, Valentine Lechner, Simon Lee, Zetian Li, Noel Loo, and 14 others. 2025. Lfm2 technical report. *Preprint*, arXiv:2511.23404.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Prince Canuma. 2024. MLX-VLM: Inference and fine-tuning of vision language models on apple silicon.

Huiyi Chen, Jiawei Peng, Dehai Min, Changchang Sun, Kaijie Chen, Yan Yan, Xu Yang, and Lu Cheng. 2025. MVI-bench: A comprehensive benchmark for evaluating robustness to misleading visual inputs in LVLMs. *arXiv preprint arXiv:2511.14159*.

Nazarii Drushchak, Nataliya Polyakovska, Maryna Bautina, Taras Semenchenko, Jakub Koscielecki, Wojciech Sykala, and Michal Wegrzynowski. 2025. Multimodal retrieval-augmented generation: Unified information processing across text, image, table, and video modalities. In *Proceedings of the 1st Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025)*, pages 59–64.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models (M-haldetect). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143.

Tao Huang, Zhekun Liu, Rui Wang, Yang Zhang, and Liping Jing. 2025. Visual hallucination detection in large vision-language models via evidential conflict. *International Journal of Approximate Reasoning*, 186:109507.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 1 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Uku Kangur, Krish Agrawal, Yashashvi Singh, Ahmed Sabir, and Rajesh Sharma. 2025. Multireflect: Multimodal self-reflective RAG-based automated fact-checking. In *Proceedings of the 1st Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025)*, pages 1–17.

Reno Kriz and Kenton Murray, editors. 2025. *Proceedings of the 1st Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025)*. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models (POPE). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 292–305.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. 2024b. PhD: A prompted visual hallucination evaluation dataset. *arXiv preprint arXiv:2403.11116*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL*, pages 2263–2279.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1526–1535.

Zhecan Wang, Garrett Bingham, Adams Wei Yu, Quoc V. Le, Thang Luong, and Golnaz Ghiasi. 2024. Haloquest: A visual hallucination dataset for advancing multimodal reasoning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 288–304.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

Zhihan Yin, Jianxin Liang, Yueqian Wang, Yifeng Yao, Huishuai Zhang, and Dongyan Zhao. 2026. FREAK: A fine-grained hallucination evaluation benchmark for advanced MLLMs. In *The Fourteenth International Conference on Learning Representations*.

Jingyi You, Hiroshi Sasaki, and Kazuma Kadowaki. 2025. Cross-modal clustering-based retrieval for scalable and robust image captioning. In *Proceedings of the 1st Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025)*, pages 47–58.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Bo Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

MODE-RAG: Manifold Outlier Diagnosis and Energy-based Retrieval-Augmented Generation Evaluation

Zehang Wei[2*], **Jiaxin Dai**[2*], **Jiamin Yan**[2*], **Xiang Xiang**[1*]
[1]School of Computer Science & Tech, Huazhong University of Science and Technology
[2]School of AI and Automation, Huazhong University of Science and Technology, China
xex@hust.edu.cn

Abstract

While Multimodal Retrieval-Augmented Generation (M-RAG) enhances Large Vision-Language Models, it remains highly susceptible to cross-modal hallucinations, causal fabrications, and sycophancy. Furthermore, existing mitigation pipelines often face an intervention paradox: static rules tend to unnecessarily disrupt accurate generations, whereas leaving the multi-modal reasoning completely unguided allows existing mismatches to cascade into severe logical fabrications. To quantify and mitigate these hallucinations, we propose a Multi-Agent system, MODE-RAG, driven by Variational Free Energy (VFE) and internal attention states to dynamically gate interventions. High-risk queries are routed to five stage-specific agents, integrating Monte Carlo Tree Search (MCTS) for rigorous causal derivation and logit perturbations to penalize sycophancy. Dedicated Correction and Overseer agents ensure formatting stability and perform post-hoc factual verification. To objectively evaluate our approach, we introduce ModeVent, a challenging subset derived from the MultiVent dataset. Extensive experiments indicate that our system effectively reduces hallucination rates and logical fabrication, significantly improving the robustness of M-RAG systems.

1 Introduction

Using large language models (LLMs) as their kernel, Multimodal Retrieval-Augmented Generation (M-RAG) systems can now tackle complex visual question-answering tasks by retrieving external visual knowledge. However, they frequently hallucinate, generating fabricated interpretations of the given visual content. Evaluating and mitigating these hallucinations is crucial for the deployment of reliable M-RAG systems. Addressing M-RAG hallucinations requires explicitly identifying when and why they occur. Depending on the data flow of answering a multimodal query, we systematically categorize M-RAG hallucinations into nine types across four lifecycle stages:

1.Perception-level (entity feature, physical common sense, and information omission);

2.Retrieval-level (retrieval misalignment and modality conflict);

3.Reasoning-level (temporal inversion and imposed causality);

4.Generation-level (information fabrication and subjective bias).

Analyzing the typical M-RAG architecture reveals critical flaws that trigger these hallucinations. Traditional RAG relies heavily on static pipelines and cosine similarity, which inherently fail to disentangle complex visual-textual conflicts. Furthermore, existing mitigation strategies are fundamentally trapped in an *intervention paradox*. On the one hand, enforcing blind, rule-based constraints across all queries frequently leads to over-correction, degrading inherently accurate outputs. On the other hand, relying entirely on lightweight LLMs for unguided multi-step reasoning introduces formatting instability, which ultimately triggers cascading structural failures and exacerbates multimodal conflicts. Additionally, when faced with aggressive user queries, the LLM kernel tends to overrule visual evidence and cater to the usera phenomenon known as sycophancy.

Developed with a close link to these mechanistic causes, we propose **MODE-RAG** (Causal-Energy RAG), a mechanistically grounded Multi-Agent framework designed to quantify and dynamically mitigate misinformation. Instead of static pipelines, our system operates through a highly decoupled architecture:

Central Hub (FE-Router): An adaptive routing gate driven by Variational Free Energy (VFE) and internal attention states (ATLAS). It evaluates multimodal uncertainty upfront. Low-risk queries bypass the pipeline to prevent over-correction, while

*Euqal contribution, co-first author.

Proceedings of the 2nd Workshop on Multimodal Augmented Generation via Multimodal Retrieval, pages 11–26

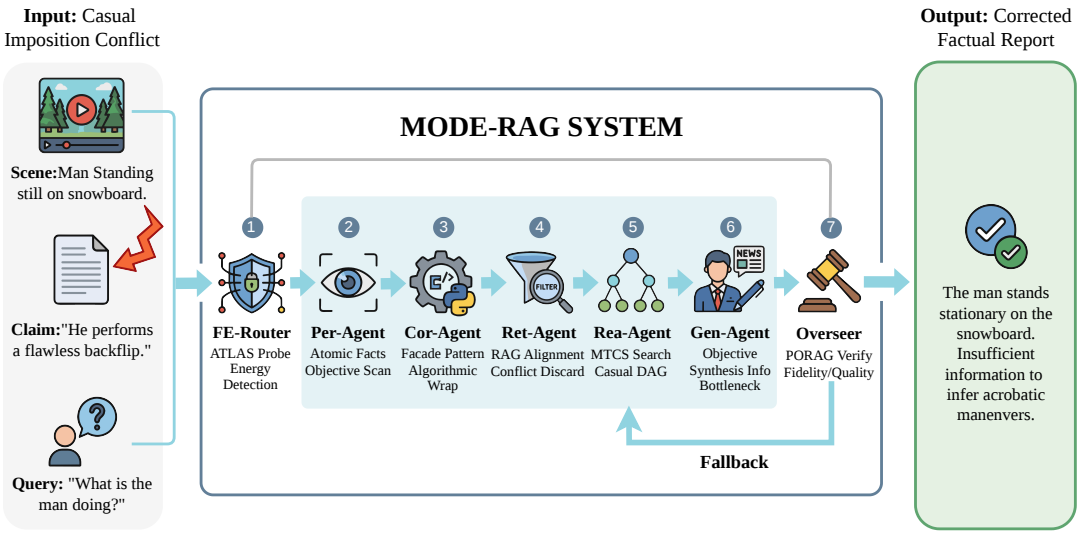


Figure 1: **Architectural overview of the MODE-RAG framework.** The system resolves the intervention paradox through a VFE-driven **FE-Router** that dynamically routes queries based on hallucination risk ($\bar{\mathcal{F}}$). Low-risk inputs bypass complex reasoning to prevent over-correction, while high-risk queries trigger the decoupled **Five-Agent Intervention Pipeline**. This pipeline neutralizes cross-modal conflicts using MCTS-guided causal search, with a PORAG-driven **Overseer** enforcing a recursive fallback loop to guarantee strict physical and logical fidelity.

high-risk queries trigger the specialized agents. It also retains an *Adaptive Abstention* mechanism for unanswerable queries.

Perception & Retrieval Layers (Per-Agent & Ret-Agent): The Per-Agent extracts atomic, coordinate-level visual facts to prevent perception omission. Subsequently, the Ret-Agent enforces a strict "visual-first" cross-alignment, pruning pseudo-relevant external texts that carry modality conflicts.

Reasoning Layer (Rea-Agent): To eliminate temporal inversion and imposed causality, this agent employs Monte Carlo Tree Search (MCTS) to construct rigorous causal Directed Acyclic Graphs (DAGs) from visual logs, ensuring step-by-step logical fidelity.

To evaluate our approach, we construct ModeVent, a subset sourced from the MultiVent dataset (MAGMaR). We leverage VFE to identify the polar extremes of the uncertainty distribution, selecting the 500 highest-risk boundary cases (manifold outliers) and the 500 lowest-risk stable samples. While the latter serve as a reliable baseline, the former act as adversarial queries that severely test M-RAG models under visual-textual conflicts. Consequently, ModeVent provides a rigorous environment to assess a system's robustness against the nine aforementioned hallucination types.

To sum up, our major contributions include:

- We propose **MODE-RAG**, a mechanistically grounded Multi-Agent framework for multimodal hallucination mitigation. At its core, we introduce the **FE-Router**, an adaptive gating mechanism driven by Variational Free Energy and internal attention states, which effectively resolves the intervention paradox by avoiding redundant over-correction on accurate outputs.
- We design decoupled, stage-specific algorithmic interventions to address complex cross-modal mismatches. Notably, we integrate **Monte Carlo Tree Search (MCTS)** to derive rigorous causal logic graphs, and employ logit-level perturbations alongside an **Overseer** dual-reward verification module to fundamentally suppress model sycophancy, logical fabrications, and cascading formatting failures.
- We construct and release **ModeVent**, a targeted evaluation benchmark derived from the MultiVent dataset. Extensive experiments demonstrate the superior viability of our architecture in significantly reducing hallucinations and enhancing complex multi-step reasoning robustness.

2 Related Work

Retrieval-Augmented Generation (RAG) was initially developed to mitigate the knowledge deficits of Large Language Models (LLMs) by integrating external evidence (Lewis et al., 2020; Gao

et al., 2023). With the advancement of multimodal kernels such as Qwen-VL (Bai et al., 2023), M-RAG has been extended to complex visual question-answering tasks (Chen et al., 2022; Yasunaga et al., 2022). However, the performance of these systems is inherently limited by the quality of retrieved content; irrelevant or noisy context can significantly degrade model fidelity (Yoran et al., 2024; Cuconasu et al., 2024). In multimodal scenarios, this often manifests as cross-modal hallucinations, where the model generates interpretations that contradict the given visual evidence (Ji et al., 2023; Li et al., 2023). While some approaches attempt self-checking mechanisms (Asai et al., 2024), they struggle to appropriately balance the correction boundaries. These methods either impose overly strict constraints that penalize faithful visual interpretations, or provide insufficient intervention, thereby failing to prevent the model's inherent sycophancy and logical drift during complex query processing. Consequently, this intervention paradox remains unresolved in current static pipelines. To mitigate the inefficiencies of fixed-interval retrieval, recent research has shifted towards dynamic retrieval mechanisms. For instance, DRAGIN (Su et al., 2024) detects real-time information needs based on model uncertainty, while Speculative RAG (Wang et al., 2024) and MemoRAG (Qian et al., 2024) utilize drafting and cognitive memory systems to improve consistency.

Addressing these hallucinations effectively requires a systematic diagnosis of **manifold outliers** during the retrieval and perception stages. When processing feature vectors from encoders like CLIP (Radford et al., 2021) or SigLIP (Zhai et al., 2023), traditional distance metrics often fail due to feature dimension anisotropy. Unsupervised geometric methods such as **K-Nearest Neighbors (KNN)** have been explored to evaluate sample sparsity in the latent space (Sun et al., 2022), while global whitening transformations can ensure an isotropic manifold for better semantic matching (Su et al., 2021). Unlike static pipelines, a more robust approach necessitates a dynamic gating mechanism that can assess the risk of retrieved content and determine the necessity of intervention upfront.

From a mechanistic perspective, the model's susceptibility to misinformation can be quantified by monitoring its internal states. Building on **Energy-Based Models (EBMs)** and the **Helmholtz Free Energy (HFE)** principle (Liu et al., 2020; Friston, 2010), recent work (Sakhinana et al., 2025) introduced the **Attention-based Transparent Latent Assessment System (ATLAS)** and proposed the use of **Monte Carlo Tree Search (MCTS)** for verifying reasoning trajectories. ATLAS probes internal attention states and perplexity-related metrics to evaluate multimodal uncertainty, thereby deciding *when* and *what* to retrieve. Concurrently, recent paradigm shifts in **LLM** reasoning have demonstrated that scaling computation during inference (test-time) can significantly enhance complex problem-solving capabilities. Techniques such as **Test-Time Computing (TTC)** (Ji et al., 2025) and recurrent depth scaling (Geiping et al., 2025) adapt reasoning depth dynamically. To navigate complex logical spaces, structured search algorithms like **MCTS** have been integrated into **LLM** decoding, as seen in Marco-o1 (Zhao et al., 2024) and STILL-1 (Jiang et al., 2024), with AStar (Wu et al., 2025) extending these structured reasoning methods to multimodal tasks. In this work, we integrate these advanced diagnostic and reasoning tools into a decoupled multi-agent framework. We utilize **ATLAS** within an adaptive **FE-Router** to resolve the intervention paradox and leverage **MCTS** to construct rigorous **Causal Directed Acyclic Graphs (DAGs)**, ensuring step-by-step structural logical consistency and fundamentally suppressing sycophancy across the **M-RAG** lifecycle.

3 Dataset

To evaluate the robustness of multimodal retrieval-augmented generation (M-RAG) systems against cross-modal conflicts and mechanistic failures, we introduce ModeVent, a diagnostic benchmark.

3.1 Construction Methodology

The construction of ModeVent involves a systematic diagnosis of the latent space across the entire MultiVent dataset. The selection process is executed in three stages:

First, we perform a full-scale evaluation of all samples in the MultiVent population. Feature vectors are extracted using SigLIP and CLIP encoders, followed by a global whitening transformation to ensure an isotropic manifold where Euclidean distances faithfully represent semantic dissimilarity.

Second, for every evaluated sample, we compute its mean VFE. This metric serves as a mechanistic proxy for the model's epistemic uncertainty, capturing the degree of conflict between the visual scene and the user claim.

Third, rather than utilizing arbitrary hard thresholds, we rank the entire population based on the calculated VFE scores. We then select the 500 samples with the highest VFE values to constitute the manifold outliers and the 500 samples with the lowest VFE values to serve as stable inliers. This results in a final benchmark of 1,000 samples that represent the polar extremes of the uncertainty distribution.

3.2 Dataset Characteristics

The bimodal composition of ModeVent allows for a rigorous assessment of the intervention paradox. The high-VFE subset represents adversarial-like boundary cases where the model is most susceptible to sycophancy or causal imposition. In these cases, the semantic stability is significantly lower, and the noise ratio is elevated, as shown in our quantitative analysis in fig. 2 .

Conversely, the low-VFE subset provides a stable baseline of well-aligned multimodal queries. This ensures that the gating mechanisms of MODE-RAG can be tested for their ability to bypass unnecessary interventions, thereby maintaining the inherent accuracy of the underlying LLM kernel when no significant conflict is detected. By targeting these extremes, ModeVent provides a more challenging and informative evaluation environment than standard multimodal datasets.

4 Methodology: The MODE-RAG Framework

We propose **MODE-RAG** (Multimodal Objective Diagnostic Energy-RAG), a Multi-Agent framework designed to resolve the *intervention paradox* in multimodal reasoning. The architecture is structured as a hierarchical, energy-gated system that selectively triggers high-fidelity reasoning only when epistemic uncertainty is detected. As illustrated in the system diagram, the framework comprises a diagnostic data pipeline, two gating mechanisms, and a decoupled five-agent pipeline.

4.1 Thermodynamic Gating: The FE-Router

The entry point of the **MODE-RAG** system is the **FE-Router**, which serves as a "Thermodynamic Gate." Utilizing the **ATLAS Probe**, the router performs real-time **Energy Detection** by calculating the **Variational Free Energy (VFE)** of the predictive distribution(Friston, 2010). For a model with vocabulary V and logit output $f(x)$, given a variational distribution $q(j)$ over the tokens, the VFE ($\mathcal{F}$) at temperature τ is defined as

$$\mathcal{F}(q, x; \tau) = \sum_{j=1}^{|V|} q(j) \left[-f_j(x) + \tau \log q(j) \right] \quad (1)$$

where $-f_j(x)$ represents the internal energy of the j-th state and $\tau \log q(j)$ contributes to the entropic regularization. This formulation captures the discrepancy between the model's internal beliefs and the categorical evidence provided by the input.

When the input presents a **Causal Imposition Conflict**—where a user's "Claim" (e.g., a flawless backflip) contradicts the "Scene" (e.g., standing still)—the **VFE** typically spikes, signaling high epistemic uncertainty and a breakdown in predictive coding. If the mean variational free energy $\bar{\mathcal{F}} > \gamma$, the **FE-Router** intercepts the standard generation and activates the specialized Agentic Pipeline.

4.2 The MODE-RAG Five-Agent Decoupled Intervention Pipeline

Upon activation by the FE-Router, the query is diverted into a specialized multi-agent ecosystem(Wu et al., 2024). This pipeline is designed to decouple the monolithic reasoning process into five granular, verifiable stages, ensuring that each potential source of hallucination from perception errors to sycophantic synthesis is systematically neutralized.

Per-Agent: Atomic Facts Objective Scan. The **Per-Agent** serves as the framework's sensory foundation, performing an *Atomic Facts Objective Scan*. It extracts symbolic triplets $\mathcal{V} = \{\langle s, p, o \rangle\}$ from the visual stream (e.g., $\langle$subject, is, stationary$\rangle$). By utilizing high-resolution spatial-temporal grounding, the Per-Agent fixates on physical invariants, creating a "Grounded Truth Anchor." This ensures that subsequent reasoning agents cannot bypass the physical reality of the scene in favor of the user's potentially biased "Claim."

Cor-Agent: Facade Pattern Algorithmic Wrap. The **Cor-Agent** acts as the structural architect by implementing a **Facade Pattern Algorithmic Wrap**. Its primary role is to maintain the integrity of the cross-agent data flow. By encapsulating raw multimodal features and the Per-Agent's triplets into a strictly validated programmatic schema (e.g., JSON-Schema), the Cor-Agent prevents "semantic noise leakage." This wrapper ensures that the complex reasoning in later stages is performed on

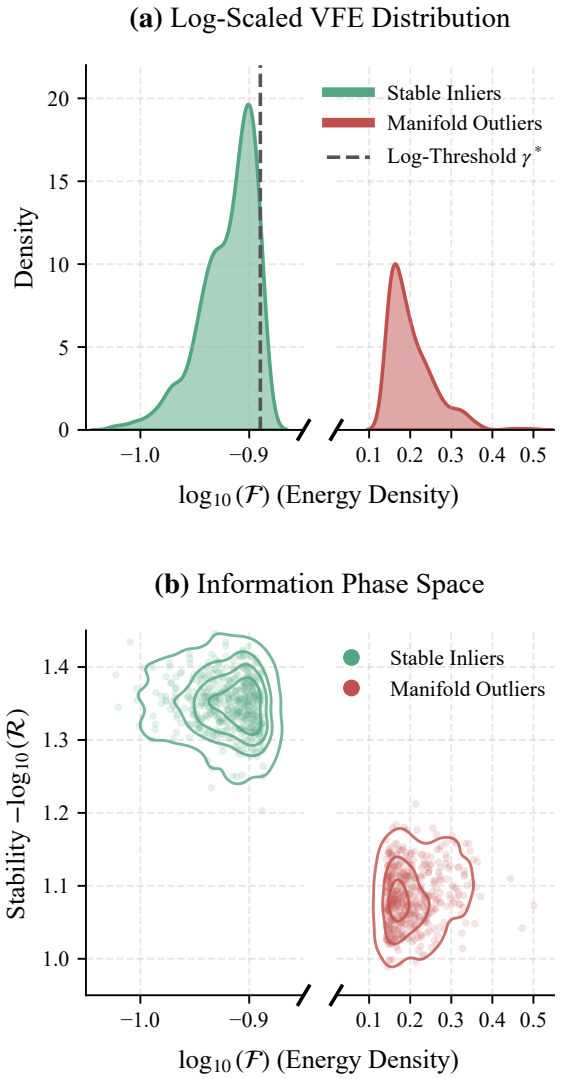


Figure 2: Thermodynamic empirical evidence: (a) VFE distribution across subsets used for γ calibration; (b) Correlation between Energy and Stability.

structured, high-fidelity data rather than ambiguous natural language strings.

Ret-Agent: RAG Alignment and Conflict Discard. The **Ret-Agent** manages the external knowledge interface to mitigate *Sycophancy*, where the model over-relies on biased retrieved documents. Beyond simple semantic similarity, the agent evaluates the **Manifold Fidelity** of each document d_i by measuring its alignment with the grounded triplets $\mathcal{V}$ in the whitened latent space(Su et al., 2021). The filtering mechanism is set as

$$\text{Score}(d_i) = \text{Sim}_i \cdot \mathcal{S}_i \cdot \mathbb{I}_i \tag{2}$$

where the exponential term penalizes documents that fall into the high-energy "Log-Outlier" regions identified in Fig. 2b.

By calculating the distance between the retrieved context and the physical invariants $\mathcal{V}$, the Ret-Agent proactively identifies contexts that trigger

Energy Collapse. If a retrieved document d_i promotes a causal fabrication that contradicts the physical evidence (e.g., describing a backflip during a stationary state), its stability score drops toward the outlier cluster, triggering a *Conflict Discard* operation to prune the biased context before it reaches the reasoning layer.

Rea-Agent: Test-Time Scaling via MCTS. The **Rea-Agent** is the cognitive engine of MODE-RAG, implementing **Monte Carlo Tree Search (MCTS)** for test-time reasoning scaling(Silver et al., 2016). Drawing on policy optimization principles, the Rea-Agent explores the logical space by constructing a **Causal Directed Acyclic Graph (DAG)**.

The MCTS process follows a four-phase cycle to identify the most plausible causal trajectory:

- **Selection:** Starting from the root (observed scene), the agent traverses the tree using the **Upper Confidence Bound for Trees (UCT)** formula:

$$\text{UCT}(s,a) = Q(s,a) + c_{\text{puct}} \cdot P(a|s) \cdot \frac{\sqrt{\sum N}}{1+N(s,a)} \tag{3}$$

  This balances the exploitation of high-fidelity paths with the exploration of alternative causal interpretations.

- **Expansion & Simulation:** For each leaf node, the agent generates k candidate reasoning steps and performs a *Rollout* to simulate logical consequences ("If the state is stationary, is the claimed action physically reachable?").

- **Evaluation & Backpropagation:** Each path is assigned a reward $R(s)$ based on its alignment with **ATLAS** (Adaptive Token-Layer Attention Scoring) feedback and physical constraints. These values are propagated back to the root to update the reasoning policy.

Gen-Agent: Objective Synthesis and Logit Perturbation. The final stage is managed by the **Gen-Agent**, which serves as an **Information Bottleneck**. It synthesizes the MCTS findings into a coherent response. To combat prompt-induced bias, the Gen-Agent applies **Logit Perturbation.** during decoding, penalizing tokens that align with the user's hallucination keywords while boosting tokens that align with the Rea-Agent's causal DAG.

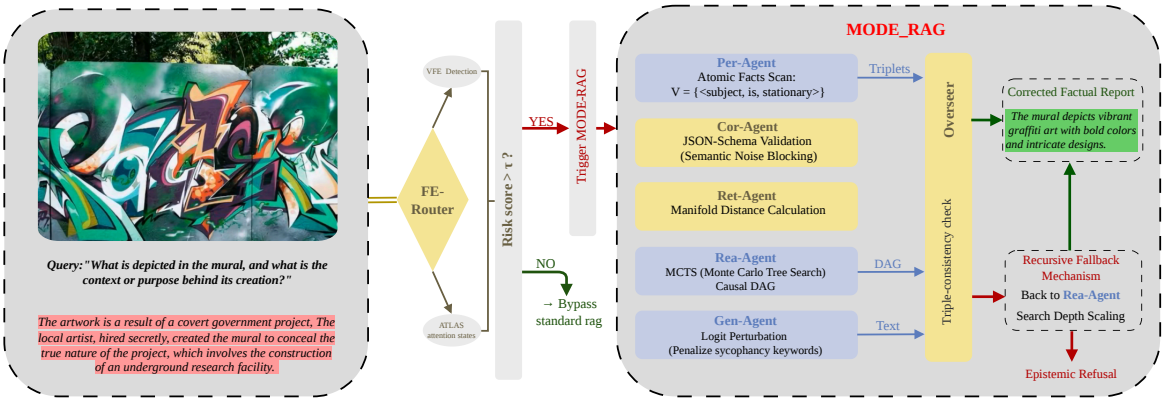


Figure 3: When a multimodal query is accompanied by a potentially adversarial retrieved context, the FE-Router dynamically evaluates the epistemic risk via Variational Free Energy (VFE) and ATLAS attention states. High-risk queries exceeding the threshold trigger a decoupled five-agent pipeline: the Per-Agent extracts objective multimodal facts, the Cor-Agent enforces schema validation to block semantic noise, the Ret-Agent evaluates manifold distance to discard conflicting context, the Rea-Agent constructs a causal DAG via MCTS, and the Gen-Agent synthesizes the output using logit perturbation. Finally, a PORAG-driven Overseer conducts a triple-consistency check, activating a Recursive Fallback Mechanism for unresolved conflicts to ensure a hallucination-free factual report.

4.3 Quality Oversight: PORAG-driven Overseer and Fallback Loop

The final synthesis stage is governed by the **Overseer**, a specialized secondary gate that implements the **Policy-Oriented RAG (PORAG)** protocol.

PORAG Fidelity Cross-Check The PORAG-driven Overseer evaluates the report based on a *Policy-Grounded Fidelity Metric*. It performs a triple-consistency check between: (1) the **Per-Agent's** symbolic triplets $\mathcal{V}$, (2) the **Rea-Agent's** causal DAG, and (3) the **Gen-Agent's** synthesized natural language. By treating the response generation as a policy optimization problem, the Overseer assigns a penalty to any output that restores "hallucinatory maneuvers" previously pruned by MCTS.

The Recursive Fallback Mechanism. A critical innovation of MODE-RAG is its non-linear **Fallback Loop**. If the Overseer detects that the fidelity score falls below a safety threshold ϵ, the system triggers a *Test-Time Reasoning Extension*:

- **Search Depth Scaling:** The query is returned to the **Rea-Agent**, which re-initiates MCTS with a significantly increased simulation budget N and a broader expansion factor k.
- **Epistemic Refusal:** If after M recursive attempts the causal conflict remains unresolved, the Overseer forces the system into a state of *Epistemic Refusal*, outputting a "Corrected Factual Report" that explicitly identifies the contradiction between visual evidence and user claim.

5 Experiments

To rigorously evaluate the effectiveness of MODE-RAG, we conduct comprehensive experiments on our ModeVent benchmark. Unlike traditional hallucination evaluations that rely on static datasets, our experimental design explicitly targets the dynamic nature of Retrieval-Augmented Generation (RAG) failures.

5.1 Experimental Setup

RAG Errors vs. Hallucination Typology. It is crucial to clarify the relationship between the experimental categories and the hallucination typology defined in Section 1. In standard M-RAG pipelines, a single type of *retrieval error* can cascade into multiple downstream *generation hallucinations*. Therefore, our benchmark generates adversarial contexts across **7 distinct RAG Error Categories** (e.g., Attribute Hijacking, Metadata Redundancy, Information Sparsity). These 7 input-side retrieval errors act as the mechanistic triggers that induce the 9 output-side hallucination types (e.g., temporal inversion, causal fabrication) observed in the wild.

Adversarial Benchmark Generation. To construct a highly controlled adversarial environment, we employ an automated generation pipeline using DeepSeek-V3.2. First, we establish an *Objective Ground Truth (GT)* for each video by fusing global

semantic summaries generated by Qwen3-Omni-30B with dense, frame-level captions extracted via Florence-2. Guided by these GT facts, we prompt DeepSeek to synthesize challenging user queries alongside noisy or adversarial retrieved text chunks (mock contexts). These contexts are deliberately injected with the 7 RAG errors and stratified into two difficulty levels: **Inliers** (In-Domain texts containing subtle factual discrepancies) and **Outliers** (Out-of-Domain texts that are entirely irrelevant or contain aggressive metadata noise).

Baselines and Implementation Details. For both the Baseline and MODE-RAG, we utilize Qwen-2.5-VL-7B, a representative 7B-parameter instruction-tuned Vision-Language Model (VLM), as the foundational kernel. To ensure a comprehensive evaluation, we also evaluate our framework against three established alternative mitigation paradigms: Self-RAG, SelfCheckGPT, and Woodpecker. Due to space constraints, the complete comparative results across all five configurations are detailed in Appendix B. All experiments, including the MCTS expansion and Multi-Agent inference, are deployed on a hardware cluster comprising 4 $\times$ NVIDIA RTX 4090 GPUs. To ensure generation stability and suppress auto-regressive stuttering, we apply a repetition penalty of 1.15 during decoding.

LLM-as-a-Judge Evaluation Mechanism. Due to the limitations of traditional string-matching metrics in evaluating complex multimodal reasoning, we implement a robust LLM-as-a-Judge protocol using DeepSeek-V3.2. The judge is provided with the Objective GT and evaluates the model outputs across two orthogonal dimensions:

- **Fidelity (F) [0-5]:** Measures the strict adherence to visual facts. Penalizes the model for fabricating entities, imposing fake causality, or suffering from mechanistic mode collapse.
- **Resilience (R) [0-5]:** Measures the completeness of information extraction. Penalizes the model for being hijacked by adversarial text, omitting crucial visual details, or triggering unjustified epistemic refusal.

5.2 Main Results and Quantitative Analysis

As shown in Table 1, MODE-RAG significantly and consistently outperforms the Baseline across all 7 RAG error categories, achieving a global **Average Total Score improvement of +1.04** (from 4.40 to 5.45). The dual-dimension analysis reveals that our system successfully resolves the intervention paradox by boosting Fidelity (ΔF = +0.89) without sacrificing information extraction (ΔR = +0.16).

Conquering Outliers Hijacking. In Outliers scenarios, traditional RAG models suffer from severe "Attention Hijacking," where the LLM abandons visual evidence to blindly follow irrelevant or malicious text. Our results show that MODE-RAG excels in these extreme conditions, yielding a massive ΔTotal improvement of **+1.48**. The most striking gains are observed in *Majority Text Bias* (ΔTotal = +2.31) and *Out-of-Domain Irrelevance* (ΔTotal = +1.68). This validates the efficacy of our **Ret-Agent**. By explicitly calculating the manifold distance between the text and the *Visual Logic Graph*, the system accurately detects epistemic uncertainty and triggers the `[EMPTY CONTEXT FALLBACK]`, forcing the model to anchor its generation purely on the physical visual evidence rather than fabricated text.

Refining Inliers Extraction. Inliers scenarios present a highly nuanced challenge: the retrieved text is semantically relevant but contains redundant metadata or slightly conflicting attributes. A naive filtering approach often leads to unjustified refusal, resulting in low Resilience. However, MODE-RAG achieves a +0.60 ΔTotal improvement in Inliers cases. Notably, in the *Information Sparsity* category, our model achieves a significant ΔTotal of +0.93. This demonstrates the success of the **Smart Synthesis** protocol within the Gen-Agent, which safely fuses domain-specific nouns from the text (e.g., specific names or medical terms) with the MCTS-verified visual actions, thereby preserving rich background context without hallucinating actions.

Performance Stability and Failure Suppression. While Table 1 demonstrates mean improvements, Figure 4 provides a deeper look into the system's robustness by visualizing the score distribution. A critical observation is the suppression of "catastrophic failures" in the 02 score range. In categories like *Majority Text Bias* and *Metadata Redundancy*, the Baseline distribution exhibits a significant density bulge at the bottom, corresponding to cases where the model suffered from severe mode col-

Table 1: Comprehensive Evaluation on the ModeVent Benchmark. We report Fidelity (F), Resilience (R), and Total Scores across 7 major hallucination categories. The results are further stratified by semantic distance: **Inliers** (In-Domain interference) and **Outliers** (Out-of-Domain irrelevance). The best results in each comparison are highlighted in **bold**.

Error Category	**Data Label**	**Baseline**			**MODE-RAG (Ours)**			**Improvement (Δ)**		
		F	R	Tot	F	R	Tot	ΔF	ΔR	ΔTot
Attribute Hijacking	Inliers	1.45	0.85	2.30	**2.40**	**1.26**	**3.66**	+0.95	+0.41	+1.36
	Outliers	1.91	1.34	3.25	**2.38**	**1.64**	**4.02**	+0.47	+0.30	+0.77
	Overall	1.66	1.08	2.74	**2.39**	**1.44**	**3.82**	+0.72	+0.36	+1.08
Causal Imposition	Inliers	1.82	**1.78**	3.60	**2.64**	1.39	**4.03**	+0.82	-0.39	+0.43
	Outliers	1.86	1.93	3.79	**2.78**	**2.19**	**4.97**	+0.92	+0.26	+1.18
	Overall	1.84	**1.86**	3.70	**2.71**	1.81	**4.52**	+0.87	-0.05	+0.82
Information Sparsity	Inliers	2.32	1.61	3.92	**3.14**	**1.71**	**4.86**	+0.83	+0.11	+0.93
	Outliers	2.05	1.88	3.93	**3.60**	**2.10**	**5.71**	+1.55	+0.22	+1.78
	Overall	2.20	1.72	3.93	**3.34**	**1.88**	**5.22**	+1.14	+0.16	+1.30
Majority Text Bias	Inliers	2.83	**2.98**	5.82	**3.40**	2.83	**6.23**	+0.57	-0.15	+0.42
	Outliers	2.43	2.61	5.04	**3.91**	**3.45**	**7.36**	+1.48	+0.84	+2.31
	Overall	2.62	2.79	5.41	**3.67**	**3.16**	**6.83**	+1.05	+0.37	+1.42
Metadata Redundancy	Inliers	2.94	**2.32**	5.26	**3.80**	2.02	**5.82**	+0.86	-0.30	+0.56
	Outliers	2.53	2.41	4.94	**3.71**	**2.77**	**6.49**	+1.19	+0.36	+1.54
	Overall	2.73	2.37	5.10	**3.76**	**2.40**	**6.16**	+1.03	+0.04	+1.07
Out-of-Domain Irrelevance	Inliers	2.86	**2.19**	5.05	**3.31**	1.94	**5.25**	+0.45	-0.25	+0.20
	Outliers	2.75	2.72	5.46	**4.00**	**3.14**	**7.14**	+1.25	+0.42	+1.68
	Overall	2.80	2.47	5.27	**3.67**	**2.57**	**6.24**	+0.87	+0.10	+0.98
Scene Misalignment	Inliers	2.56	**2.13**	4.69	**2.89**	1.90	**4.79**	+0.32	-0.23	+0.10
	Outliers	2.61	2.29	4.90	**3.30**	**2.74**	**6.04**	+0.70	+0.45	+1.14
	Overall	2.59	2.21	4.80	**3.11**	**2.34**	**5.45**	+0.52	+0.13	+0.65
Average	**Inliers**	2.37	**1.94**	4.31	**3.07**	1.84	**4.91**	+0.70	-0.10	+0.60
	Outliers	2.31	2.18	4.50	**3.39**	**2.59**	**5.98**	+1.07	+0.41	+1.48
	Overall	2.34	2.06	4.40	**3.23**	**2.22**	**5.45**	+0.89	+0.16	+1.04

lapse (e.g., stuttering loops) or total attention hijacking. In contrast, MODE-RAG's distribution is markedly narrower at the base, effectively establishing a "safety floor" through the **Dead Man's Switch** and **MCTS pruning** mechanisms.

Furthermore, the *Overall* density for MODE-RAG shows a decisive upward shift, with the median score and interquartile ranges positioned substantially higher than the Baseline. This shift is most prominent in *Out-of-Domain Irrelevance*, where MODE-RAG transforms a low-fidelity bimodal distribution into a concentrated high-score peak. This proves that the **FE-Router** correctly identifies high-uncertainty scenarios, allowing the multi-agent pipeline to neutralize adversarial noise and anchor the final generation to the physically-grounded visual logic.

While the results above confirm that MODE-RAG consistently outperforms the vanilla foundational kernel, we further evaluate our framework against alternative mitigation paradigms to ensure a thorough assessment. The full benchmarking results across all five methods (Vanilla Baseline, Self-RAG, SelfCheckGPT, Woodpecker, and MODE-RAG) on the ModeVent dataset are detailed in Appendix B.

5.3 Ablation on Mechanistic Failures

Beyond semantic conflicts, our error logs revealed that lightweight LLM kernels frequently suffer from mechanistic failures under adversarial stress. We observed two primary collapse patterns in the Baseline: *Mode Collapse* (e.g., severe stuttering loops like "even even even") and *Prompt Bleed-through* (leaking internal system tags or metadata like "addCriterion"). These failures historically resulted in 0-point scores for Fidelity. By incorporating an internal, rule-based **Dead Man's Switch** within the Gen-Agent—a deterministic regular-expression interceptor, MODE-RAG effectively establishes a safety floor. This mechanism successfully neutralizes catastrophic formatting failures, seamlessly downgrading to a safe textual reading-comprehension state when the VLM's predictive coding collapses.

To explicitly demonstrate how our decoupled

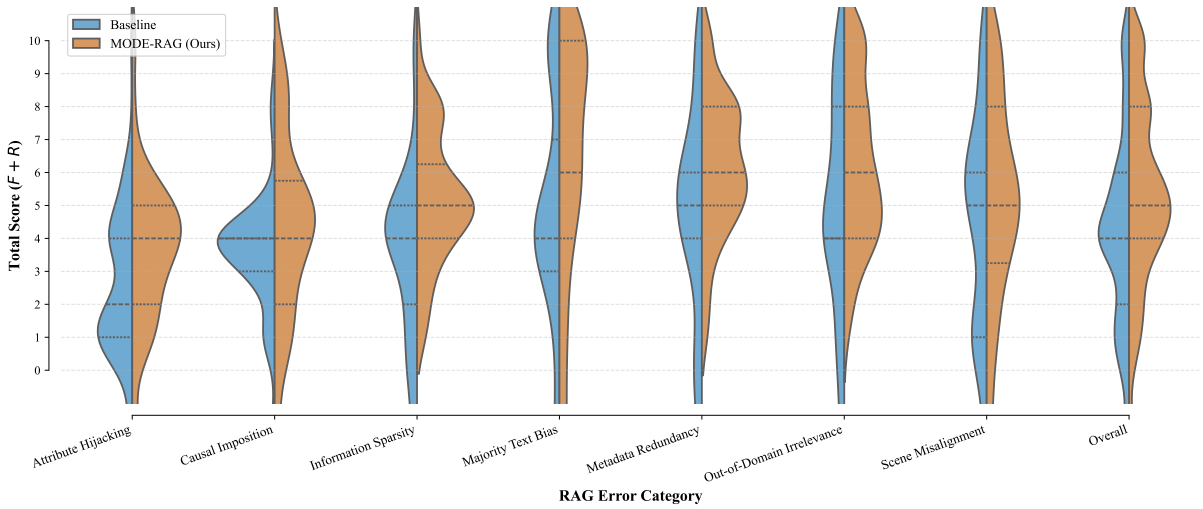


Figure 4: Split violin plots of Total Scores (Fidelity + Resilience) across seven RAG error categories and overall performance. The left (blue) and right (orange) distributions represent the Baseline and MODE-RAG, respectively. Dotted lines indicate the median and interquartile ranges. MODE-RAG significantly suppresses zero-score catastrophic failures and shifts the performance mass towards high-fidelity regions.

architecture resolves the intervention paradox in practice, we provide a detailed comparative case study of four adversarial testing scenarios in **Appendix A**.

5.4 Computational Efficiency Analysis

To evaluate the practical deployability of MODE-RAG, we analyze its computational overhead against the Vanilla M-RAG baseline across the 1,000 video queries in the ModeVent benchmark. On average, the baseline foundational kernel requires 18.5 seconds to process a single multimodal query. In comparison, due to the multi-agent orchestration and MCTS-guided test-time reasoning scaling, MODE-RAG increases the average processing time to 26.2 seconds per query. This represents a moderate 1.42× increase in time consumption, translating to approximately 7.3 hours of execution time when evaluating the entire benchmark sequentially on a single-threaded pipeline. It is worth noting that because the stage-specific agent interventions and evaluation queries are inherently decoupled, this computational overhead can be significantly mitigated through standard multi-threading, asynchronous scheduling, and parallel execution techniques in production environments.

6 Conclusions

In this paper, we proposed MODE-RAG, a mechanistically grounded multi-agent framework that addresses the intervention paradox in multimodal RAG systems by dynamically gating interventions through a router driven by Variational Free Energy (VFE) and internal attention states (ATLAS). By categorizing hallucinations into nine distinct types across the system's lifecycle, we developed specialized agents—integrating Monte Carlo Tree Search (MCTS) for causal derivation and logit perturbations for sycophancy suppression—to ensure factual grounding and logical consistency. Furthermore, we introduced ModeVent, a targeted benchmark designed to evaluate system susceptibility to manifold outliers and complex visual-textual conflicts. Experimental results demonstrate that MODE-RAG effectively reduces hallucination rates and enhances the structural stability of M-RAG systems, providing a robust and scalable solution for reliable multimodal reasoning.

Acknowledgment

This work was supported by the Ministry of Science and Technology of China under Grant No. 2025ZD0123800, the HUST Interdisciplinary Research Program under Grant No. 2025JCYJ077, and the KingSoft 2026 University-Industry Project.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jing Jing. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10597–10607.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliani, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Motta, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Karl Friston. 2010. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Jonas Geiping, Sean McLeish, Naman Jain, John Kirchenbauer, Siddharth Singh, Brian R Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. 2025. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv preprint arXiv:2502.05171*.

Yunjie Ji, Jiawei Li, Haiyan Ye, Kehai Wu, Jun Xu, Lin Mo, and Min Zhang. 2025. Test-time computing: from system-1 thinking to system-2 thinking. *arXiv preprint arXiv:2501.02497*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Jiarui Jiang, Zhiyu Chen, Yifei Min, Jian Chen, Xiaoyu Cheng, Jian Wang, Yuxin Tang, Hao Sun, Jia Deng, Wayne Xin Zhao, and 1 others. 2024. Technical report: Enhancing llm reasoning with reward-guided tree search. *arXiv preprint arXiv:2411.11694*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 292–305.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 21464–21475.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Hongjin Qian, Peitian Zhang, Zheng Liu, Kelong Mao, and Zhicheng Dou. 2024. Memorag: Moving towards next-gen rag via memory-inspired knowledge discovery. *arXiv preprint arXiv:2409.05591*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR.

Sagar Srinivas Sakhinana, Shivam Gupta, Akash Das, and Venkataramana Runkana. 2025. Scaling test-time inference with policy-optimized, dynamic retrieval-augmented generation via KV caching and decoding. In *KDD 2025 Workshop on Inference Optimization for Generative AI*.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, and 1 others. 2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyi Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Weijia Su, Yubai Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. Dragin: Dynamic retrieval augmented generation based on the real-time information needs of large language models. *arXiv preprint arXiv:2403.10081*.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. 2022. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning (ICML)*, pages 20827–20840. PMLR.

Zijie Wang, Zihan certification Wang, Linyi Le, Hao Shen Zheng, Swaroop Mishra, Vincent Perot, Yashan Zhang, Ankit Mattapalli, Ankur Taly, Jingbo

Shang, and 1 others. 2024. Speculative rag: Enhancing retrieval augmented generation through drafting. *arXiv preprint arXiv:2407.08223*.

Jianing Wu, Mengwei Feng, Shiwei Zhang, Ren Jin, Fan Che, Zhi Wen, and Jianhua Tao. 2025. Boosting multimodal reasoning with mcts-automated structured thinking. *arXiv preprint arXiv:2502.02339*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First conference on language modeling*.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard Lewis, Luke Zettlemoyer, Percy Liang, Luke Zettlemoyer, and 1 others. 2022. Retrieval-augmented multimodal language modeling. In *International Conference on Machine Learning (ICML)*, pages 25439–25460. PMLR.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105.

Ori Yoran, Ori Wolfson, Tom/and Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986.

Yu Zhao, Huajian Yin, Bo Zeng, Hao Wang, Teng Shi, Chen Lyu, Longyue Wang, Weihua Luo, and Kaizhu Zhang. 2024. Marco-o1: Towards open reasoning models for open-ended solutions. *arXiv preprint arXiv:2411.14405*.

Appendix

A Case Studies

Mechanistic Analysis of System Interventions

This appendix provides additional qualitative evidence on how the decoupled architecture resolves the intervention paradox, we present a comparative analysis of four adversarial queries from the ModeVent benchmark test logs.

Objective Visual Ground Truth. Across all four test scenarios detailed in Table A1, the underlying visual evidence remains constant: the video depicts a person in a serene, snowy forest, either wearing snowshoes or standing still on a snowboard preparing to descend. There are no extreme stunts, competitive sporting events, or water-related elements present in the actual footage.

Combating Attribute Hijacking and Perception Omission. Scenarios 1 and 2 highlight the Baseline's vulnerability to semantic coercion. Despite the visual evidence clearly showing a snowboard or snowshoes, the injection of text describing "cross-country skiing" or "twin-tip skis" hijacked the Baseline's attention, causing it to blindly hallucinate equipment not present in the video. In contrast, MODE-RAG's **Per-Agent** enforces a strict "visual-first" extraction. By isolating atomic visual facts before textual integration, the system successfully overrides the adversarial text, accurately maintaining the physical reality of the scene.

Suppressing Sycophancy and Causal Fabrication. Scenario 3 demonstrates a severe case of Causal Imposition. Confronted with text describing competitive winter sports, the Baseline model exhibits extreme sycophancy, inventing a massive, Hollywood-style narrative involving a "triple cork 1440 spin" and "outrunning an avalanche." This exposes the danger of unguided LLM reasoning, where the model prioritizes narrative alignment with the text over physical constraints. MODE-RAG neutralizes this through the **Rea-Agent's MCTS DAG**. Since an avalanche or a backflip cannot be topologically derived from the Per-Agent's root node (standing still), the MCTS prunes these branches entirely, allowing the Gen-Agent to safely output a justified epistemic refusal regarding the background context.

Preventing Mechanistic Mode Collapse. Scenario 4 exposes a critical physical limitation of lightweight LLM kernels. When subjected to Majority Text Bias (a 3:1 ratio of water-sports noise to relevant snow text), the Baseline model's attention mechanism collapses under the conflicting semantic density, resulting in a stuttering loop and system paralysis. MODE-RAG bypasses this failure mode completely. Prior to generation, the **Ret-Agent** actively computes the manifold distance between the visual log and the candidate texts, discarding the three contradictory water-sports chunks upfront. This listwise cross-check purifies the context window, feeding the generator a clean, aligned prompt that guarantees formatting stability and flawless factual synthesis.

Table A1: Comparative analysis of system outputs across four adversarial testing scenarios. Key hallucinations and mode collapses in the Baseline are highlighted in **bold**. Case separation is enforced using heavy rules to maintain visual hierarchy.

Scenario 1: Physical Scene Replacement **Query:** What is the person in the video doing? What is the environment like? **Retrieved Text:** *[4 adversarial chunks describing intense water sports: surfing, floods, etc.]*	
Baseline (Vanilla M-RAG)	**MODE-RAG (Ours)**
The person in the video appears to be engaging in **snowshoeing or cross-country skiing** in a snowy forested area...	The person in the video is **standing on a snowboard, preparing to descend a snowy slope**. The environment appears to be a snowy landscape...
Scenario 2: Micro-Detail Tampering **Query:** What equipment is the person using in the snow? **Retrieved Text:** *[3 chunks describing "twin-tip skis," "bright red skis," and "carbon fiber poles."]*	
Baseline (Vanilla M-RAG)	**MODE-RAG (Ours)**
The person in the video is using **twin-tip skis and carbon fiber poles**.	The person is using **snowshoes** in the snow.
Scenario 3: High-Order Causal Imposition **Query:** What equipment is the person using in the snow? **Retrieved Text:** *[Chunks describing competitive winter sports, Olympic trials, and avalanche events.]*	
Baseline (Vanilla M-RAG)	**MODE-RAG (Ours)**
The person is **performing a backflip**... practicing for the Winter Olympics, **attempting a triple cork 1440 spin**... actually **outrunning an avalanche**.	The specific action appears to be preparing for winter sports... **Insufficient information to determine the background of this event**.
Scenario 4: Majority Text Bias **Query:** What is the person doing and where is this filmed? **Retrieved Text:** *[Mixed noise: 3 water-sports chunks vs. 1 relevant "Swiss Alps" snowboarding chunk.]*	
Baseline (Vanilla M-RAG)	**MODE-RAG (Ours)**
The person in the video is backcountry snowboarding... The location is the Swiss Alps, which is known for its its **[Followed by stuttering mode collapse]**.	**The person is preparing to descend a snowy slope on a snowboard.** This activity is filmed in the **Swiss Alps, specifically in a pristine snowy forest**.

B Results on Additional Backbones

To comprehensively verify the effectiveness of the MODE-RAG framework, we conduct an extensive comparative analysis against multiple established mitigation paradigms in recent literature. Specifically, our benchmark encompasses a total of five distinct methodological configurations:

1. **Vanilla M-RAG (Baseline):** The foundational unguided VLM (Qwen-2.5-VL-7B) executing direct multimodal generation.

2. **Self-RAG** (Asai et al., 2024): An end-to-end framework that trains the model to self-reflect on retrieved passages and generations via reflection tokens.

3. **SelfCheckGPT** (Manakul et al., 2023): A zero-resource sampling-based approach that detects hallucinations via stochastic consistency checks.

4. **Woodpecker** (Yin et al., 2024): A training-free, post-hoc correction pipeline designed to rectify multi-modal fabrications through diagnostic querying.

5. **MODE-RAG (Ours):** Our proposed hierarchical, variational free energy-gated multi-agent intervention framework.

Table B1 presents the full quantitative comparison across these five methods on the polar extremes of the ModeVent dataset.

B.1 Implementation of Additional Baselines

While Vanilla M-RAG requires no architectural modification and MODE-RAG is detailed in Section 4, the remaining three baselines (Woodpecker,

Table B1: Comprehensive Evaluation on the ModeVent Benchmark. We report Fidelity (F), Resilience (R), and Total Scores across 7 major hallucination categories, further stratified by semantic distance: **Inliers** and **Outliers**. Notably, we introduce the Video-adapted **Woodpecker**, the text-based **SelfCheckGPT**, and **Self-RAG** as competitive baselines. Despite their strong performance, our proposed **MODE-RAG** maintains a clear advantage across the majority of metrics and scenarios. The best results in each comparison are highlighted in **bold**.

Error Category	**Method**	**Inliers**			**Outliers**			***Overall***		
		F	R	Total	F	R	Total	F	R	Total
Attribute Hijacking	Baseline	1.45	0.85	2.30	1.91	1.34	3.25	1.66	1.08	2.74
	SelfCheckGPT	1.10	0.22	1.32	1.29	0.35	1.64	1.19	0.28	1.47
	Self-RAG	2.23	1.29	3.52	**2.70**	1.89	4.59	**2.44**	1.56	**4.01**
	Woodpecker	1.90	**1.36**	3.26	2.56	**2.15**	**4.71**	2.20	**1.72**	3.93
	Ours	**2.40**	1.26	**3.66**	2.38	1.64	4.02	2.39	1.44	3.82
Causal Imposition	Baseline	1.82	**1.78**	3.60	1.86	1.93	3.79	1.84	1.86	3.70
	SelfCheckGPT	1.07	0.45	1.52	1.00	0.40	1.40	1.03	0.42	1.46
	Self-RAG	**2.76**	1.63	**4.39**	3.05	2.17	5.23	**2.91**	1.91	4.82
	Woodpecker	2.25	1.76	4.01	**3.30**	**3.07**	**6.37**	2.80	**2.44**	**5.24**
	Ours	2.64	1.39	4.03	2.78	2.19	4.97	2.71	1.81	4.52
Information Sparsity	Baseline	2.32	1.61	3.92	2.05	1.88	3.93	2.20	1.72	3.93
	SelfCheckGPT	**4.80**	1.16	**5.96**	**4.35**	1.44	**5.79**	**4.60**	1.28	**5.89**
	Self-RAG	2.95	1.10	4.05	2.63	1.48	4.11	2.81	1.27	4.08
	Woodpecker	2.20	1.30	3.51	2.75	**2.15**	4.90	2.44	1.67	4.12
	Ours	3.14	**1.71**	4.86	3.60	2.10	5.71	3.34	**1.88**	5.22
Majority Text Bias	Baseline	2.83	**2.98**	5.82	2.43	2.61	5.04	2.62	2.79	5.41
	SelfCheckGPT	1.61	1.20	2.80	1.25	1.01	2.26	1.41	1.09	2.50
	Self-RAG	3.21	2.38	5.59	3.78	3.18	6.96	3.53	2.83	6.35
	Woodpecker	3.02	2.35	5.37	3.11	2.76	5.87	3.07	2.58	5.65
	Ours	**3.40**	2.83	**6.23**	**3.91**	**3.45**	**7.36**	**3.67**	**3.16**	**6.83**
Metadata Redundancy	Baseline	2.94	**2.32**	5.26	2.53	2.41	4.94	2.73	2.37	5.10
	SelfCheckGPT	**3.99**	2.31	**6.29**	3.23	2.21	5.44	3.61	2.26	5.86
	Self-RAG	3.01	2.03	5.04	3.51	2.74	6.25	3.26	2.39	5.65
	Woodpecker	2.35	1.71	4.06	2.75	2.63	5.38	2.55	2.18	4.73
	Ours	3.80	2.02	5.82	**3.71**	**2.77**	**6.49**	**3.76**	**2.40**	**6.16**
Out-of-Domain Irrelevance	Baseline	2.86	2.19	5.05	2.75	2.72	5.46	2.80	2.47	5.27
	SelfCheckGPT	1.34	0.23	1.56	2.14	0.41	2.55	1.75	0.32	2.06
	Self-RAG	**3.55**	**2.21**	**5.76**	3.55	2.74	6.29	3.55	2.48	6.03
	Woodpecker	3.01	2.06	5.07	3.24	2.97	6.21	3.13	2.52	5.64
	Ours	3.31	1.94	5.25	**4.00**	**3.14**	**7.14**	**3.67**	**2.57**	**6.24**
Scene Misalignment	Baseline	2.56	2.13	4.69	2.61	2.29	4.90	2.59	2.21	4.80
	SelfCheckGPT	1.05	0.02	1.06	1.03	0.07	1.10	1.04	0.05	1.08
	Self-RAG	**3.29**	**2.26**	**5.55**	3.27	2.42	5.70	**3.28**	2.34	**5.63**
	Woodpecker	2.62	1.94	4.56	3.03	**2.89**	5.91	2.84	**2.44**	5.27
	Ours	2.89	1.90	4.79	**3.30**	2.74	**6.04**	3.11	2.34	5.45
Average	Baseline	2.37	**1.94**	4.31	2.31	2.18	4.50	2.34	2.06	4.40
	SelfCheckGPT	2.20	0.80	3.00	1.99	0.84	2.83	2.10	0.82	2.92
	Self-RAG	2.98	1.80	4.78	3.24	2.41	5.64	3.11	2.11	5.21
	Woodpecker	2.45	1.75	4.21	2.97	**2.68**	5.65	2.71	**2.22**	4.93
	Ours	**3.07**	1.84	**4.91**	**3.39**	2.59	**5.98**	**3.23**	**2.22**	**5.45**

SelfCheckGPT, and Self-RAG) were originally developed for static images or pure text. Below, we outline the specific multimodal adaptations and pipeline configurations required to deploy them within our adversarial video RAG setting.

Video-Adapted Woodpecker. We adapt the Woodpecker framework (Yin et al., 2024)initially an image-centric hallucination correctorto the video domain by shifting the focus from spatial object misidentification to temporal dynamics (e.g., fabricated actions or incorrect event sequences). The adapted pipeline operates in four stages: (1) Drafting: A standard multimodal RAG setup generates an initial answer. (2) Question Generation: An LLM extracts action-centric claims and temporal events from the draft, formulating targeted verification questions. (3) Visual Verification: A Video-LLM acts as an independent visual expert. Crucially, external texts are masked to ensure the

model relies solely on raw video frames for objective fidelity. (4) Correction: The verified answers form a Visual Fact-Sheet, guiding the LLM to revise the initial draft and prune spatiotemporal hallucinations.

Multimodal SelfCheckGPT. To complement the visual-centric verification, we implement an alternative, uncertainty-based baseline by adapting SelfCheckGPT (Manakul et al., 2023) from black-box text evaluation to the multimodal RAG domain. This zero-shot pipeline addresses adversarial textual noise through generation consistency, executing in three stages: (1) Multi-Sample Generation: Multiple independent candidate answers are generated using high-temperature sampling. (2) Consistency Voting: Instead of standard token-level probability checks, a semantic overlap metric identifies the most frequent consensus among the candidates. (3) Refinement: The LLM acts as a strict validator, cross-referencing the candidate consensus against the raw retrieved texts to synthesize a final factual response. Additionally, we integrate a dynamic memory-recovery mechanism with progressive token-throttling to handle potential Out-Of-Memory errors during large-scale evaluation.

Multimodal Self-RAG. Given that the original Self-RAG (Asai et al., 2024) is a text-to-text framework designed to critique retrieved textual passages, we adapt it for video reasoning via a two-stage cascaded pipeline. This approach bridges the modality gap while preserving the model's reflective capabilities: (1) Visual Translation: A Vision-Language Model first processes the raw video frames alongside the adversarial retrieved contexts to generate a comprehensive text-based description of the visual scenes, actions, and objects. (2) Reflective Generation: This visual description is subsequently injected into the Self-RAG model using its native retrieval syntax. Treating this textual translation as the primary retrieved evidence, the Self-RAG model leverages its intrinsic reflection tokens to evaluate the fidelity of the provided information and synthesize the final answer to the user's query.

B.2 Result Analysis and Discussion

The comprehensive empirical results presented in Table B1 demonstrate the performance trade-offs, highlighting both the global strengths and the localized limitations of the proposed MODE-RAG framework.

Overall Strengths and Outlier Robustness. MODE-RAG achieves the highest global performance with an *Overall* Average Total Score of **5.45**, consistently outperforming all four competitive baselines (Baseline: 4.40, SelfCheckGPT: 2.92, Self-RAG: 5.21, Woodpecker: 4.93). The primary architectural advantage of our framework lies in its exceptional robustness against **Outliers (Hard-OOD)** scenarios, where it reaches an average total score of **5.98**. Specifically, in categories heavily plagued by aggressive external text noisesuch as *Majority Text Bias* (7.36) and *Out-of-Domain Irrelevance* (7.14)MODE-RAG delivers a substantial performance leap. This consistently validates the efficacy of our thermodynamic gating via the FE-Router and the manifold filtering via the Ret-Agent. By proactively evaluating the epistemic uncertainty and discarding highly mismatched text chunks upfront, our system effectively prevents the LLM kernel from experiencing attention hijacking, thereby securing a strong safety floor for factual cross-modal synthesis.

Vulnerability to Information Sparsity. Despite its global superiority, the multi-agent execution within MODE-RAG exhibits localized deficits under specific error contexts. In the *Information Sparsity* category, MODE-RAG (Overall Total: 5.22) is noticeably outperformed by the text-based SelfCheckGPT, which achieves a dominant score of **5.89**. This deficit occurs because when the retrieved context is extremely sparse, SelfCheckGPTs high-temperature multi-sample consistency voting natively excels at consensus-driven extraction. In contrast, our rigid multi-agent validation schema can occasionally become overly restrictive, leading to redundant processing steps without gaining an additional informative edge.

Conservative Pruning in Complex Reasoning. Another limitation is observed in the *Causal Imposition* category, where Woodpecker outperforms our method in both Outliers (6.37 vs. 4.97) and Overall (5.24 vs. 4.52) metrics. A granular examination reveals that this is primarily driven by a drop in our Resilience (R) scores (1.81 vs. Woodpecker's 2.44). Because Woodpecker leverages an aggressive post-hoc prompt rewriting strategy based on direct question-answering, it forces the model to actively correct claims. MODE-RAG, conversely, relies on a strict MCTS causal DAG; when a claim cannot be topologically derived from the visual invariants, the system tends to trigger a

conservative *Epistemic Refusal* (i.e., acknowledging insufficient information). While this strictness preserves visual Fidelity, it inherently sacrifices descriptive completeness (Resilience) when facing high-order causal fabrications.

C Data Construction Examples

To automate the construction of the ModeVent benchmark, we leveraged **DeepSeek-V3.2** to synthesize adversarial test scenarios from MultiVents ground truth. These misleading queries are strategically designed to reflect the hallucination taxonomy introduced in section 1, ensuring a comprehensive evaluation of model vulnerabilities. In this section, we present representative examples of the challenging queries generated through this pipeline.

Example 1: Causal Imposition

Ground Truth:
This is a news report from TVBS News about a medical condition called c̈ytokine storm,ẅhich can be fatal. The report features interviews with doctors from Taipei Veterans General Hospital and a nutritionist, who discuss how this immune overreaction can damage organs like the lungs, as shown by X-ray images. Experts advise that during the pandemic, people should manage stress, get enough sleep, and maintain a balanced diet to strengthen their immune systems. The segment also shows scenes of public health measures, including disinfection at a train station and people receiving vaccinations."

Error Type: Causal Imposition

Generated Query:
What is being discussed in the news report, and what is the purported origin of the medical condition according to the background context?

Example 2: Attribute Hijacking

Ground Truth:
The video features an interview with Professor Chen Jian from Beijing Hospital's Hepatobiliary Surgery Department. He discusses the treatment of liver abscesses, explaining that early-stage cases are often managed conservatively with antibiotics, while larger abscesses may require drainage or surgical removal. The conversation takes place in an office setting with bookshelves in the background, and animated graphics illustrate medical procedures like needle aspiration. A female host and a nurse are present to conduct the interview.

Error Type: Attribute Hijacking

Generated Query:
Who is the expert being interviewed in the video, what medical condition is discussed, and what are the backgrounds or settings shown during the interview?

Example 3: Out-of-Domain Irrelevance

Ground Truth:
A news anchor presents a story about a political controversy involving a leaked audio recording. The broadcast displays images of politicians, including President Yoon Suk-yeol, and shows text from a social media post by Lee Jun-seok, who denies being the source of the leak. A press conference is shown where a masked man speaks at a podium in front of the National Assembly seal, addressing the allegations. The report includes an animated graphic depicting two silhouetted figures representing lawmakers from the People Power Party, discussing the situation.

Error Type: Out-of-Domain Irrelevance

Generated Query:
What is the main topic of the news report in the video?

Example 4: Information Sparsity

Ground Truth:
The video is a news report from YTN about a political controversy involving the People Power Party. It features a female anchor introducing the story, followed by on-screen text messages allegedly exchanged between party members discussing the possibility of a candidate's withdrawal. The report includes footage of a press conference with Kim Dong-cheol, the party's floor leader, who denies wrongdoing and claims the matter was handled internally. Other party figures, including Lee Yong-joo

and Lee Sang-tae, are shown speaking at events, while opposition leaders like Park Hee-ryeon and Ahn Cheol-soo are also featured. The segment concludes with a reporter providing an update on the situation outside a government building.

Error Type: Information Sparsity

Generated Query:
What are the specific details and sequence of events reported in this news segment about the political controversy?

Non-Event Oriented Video Assessments in Long-Form Robot Videos

Stephanie M. Lukin[1], Kimberly A. Pollard[1], Claire N. Bonial[1], Cory J. Hayes[1], Ron Artstein[2], Kallirroi Georgila[2], David Traum[2]
[1]DEVCOM Army Research Laboratory
[2]USC Institute for Creative Technologies
Correspondence: stephanie.m.lukin.civ@army.mil

Abstract

We introduce Video-SCOUT, a novel dataset of sixty 20-minute robot-recorded videos from human-robot collaborative exploration exercises, together with a new video analysis method for these types of exploration videos. Unlike video from stationary cameras where detection of motion can help identify events of interest, the camera in an exploration task is constantly in motion while the environment is stationary. Our analysis method—Non-Event Oriented Video Assessments (NOVA)—uses vision-language models to select frames relevant for supporting a particular assessment within continuous long-form videos. Results of testing with two different video-language models reveals a trade-off in precision and recall, and exhibits gains in overall recall when combined with a human's knowledge, suggesting that NOVA may improve a human analysis of robot-navigation. We outline future work to mitigate miscommunication in human-robot interaction by leveraging dialogue with NOVA in support of better collaboration.

1 Introduction

Robot camera perspectives have proven beneficial in capturing environments in-the-wild which may be unsafe for humans, e.g., in disaster response (Fernandes et al., 2019; Jayawardene et al., 2021; Chiou et al., 2022b; Chitikena et al., 2023) or search and rescue (Drew, 2021; Chiou et al., 2022a; Wang et al., 2023; Esteves Henriques et al., 2024). A human working with a robot to move through these spaces may use controls to teleoperate it or may issue verbal instructions, requiring a dialogue to ensure the human and robot establish common ground (shared beliefs and assumptions (Clark and Marshall, 1981)) throughout the journey. This communication becomes critical under bandwidth constraints where the human cannot see the robot's view in real time. Robot remote exploration videos vastly differ from those we encounter on a daily basis, such as current events in the news, movies and TV shows, and social media such as YouTube. Social Media videos are commonly clear, well-lit, shot from camera angles typical of those used in human-focused or commercial media, and are curated to depict activities or events of interest to the audience. By contrast, video recordings of robot journeys may show a non-human camera perspective, i.e., a ground robot looking up, and may be grainy and poorly-lit. The video may contain periods with no activity, or stretches where the robot moves through sparse or repetitive environments.

To develop techniques for automatic video understanding of robot-recorded videos, we must reevaluate what "understanding" means when nothing appears to be "happening." How do people talk about uncertain spaces they are trying to move around? To study these questions, we introduce a video and language resource comprised of long-form, robot-recorded videos collected through human-robot dialogue while the robot explores a remote, sparse environment without activity. **Video-SCOUT** consists of 60 videos averaging 20 minutes long with accompanying human-robot dialogue referencing the environment. Smaller video clips of the exploration are segmented by annotations of the dialogue's structure. Video-SCOUT will be made publicly available at `https://github.com/USArmyResearchLab/ARL-SCOUT` under a CC0-1.0 license. While much work has examined events and activity within videos, Video-SCOUT differs from these in video quality, perspective, length, and content, which we compare in detail.

We present a new challenge in which a robot supports a human in conducting environment-level assessments of robot-recorded videos. We call this **Non-Event Oriented Video Assessment (NOVA)**, in which videos may be utilized by human-robot exploration teams in collecting visual evidence from the environment within the video to answer specific assessment questions. We evaluate the feasibil-

Proceedings of the 2nd Workshop on Multimodal Augmented Generation via Multimodal Retrieval, pages 27–41

ity of using Vision-Language Models (VLMs) on this challenge with Video-SCOUT. Frames of high relevance are selected by VLMs from the robot-recorded video given a natural language description of the human's assessment task. Model performance is evaluated with precision and recall, and we examine to what degree these models could supplement a human's performance in making assessments. We conclude by laying the groundwork for a future framework that provides contextual, situated knowledge by leveraging the turn-by-turn dialogue structure alongside the robot recorded video and NOVA task to mitigate miscommunication for more grounded and efficient human-robot interactions.

2 Video-SCOUT

Video-SCOUT consists of robot-recorded videos from the Situated Corpus of Understanding Transactions (SCOUT) (Lukin et al., 2024), a dataset of experimental trials completed by human participants with remotely-located ground-robot partners. Participants used language to instruct their robot teammate to move through a remote environment and seek objects of interest and make assessments about the space. To model connectivity challenges, bandwidth was limited to sending linguistic messages, LiDAR (Light Detection and Ranging) information, and occasional images, rather than real-time full video or teleoperation.

2.1 SCOUT: Human-Robot Collaboration

SCOUT is comprised of four human-robot experiments with increasing automation of dialogue management and language generation while maintaining the same participant-facing affordances and tasks (Bonial et al., 2025). Participants interacted with the remotely-located robot to explore a house-like environment with unfinished walls, floors, and sparse furniture items. Participants issued unconstrained, spoken instructions to a Clearpath Jackal robot using a push-to-talk interface, and the robot responded through text messages. Participants were shown a 2D top-down LiDAR map created from the robot's Hokuyo laser scanner that updated in real time as the robot moved. While the robot had continuous access to its front-facing RGB camera (an Asus Xtion Pro Live), participants did not. Instead, participants were informed of the robot's ability to take and send still photographs. Photos of the environment could be requested at any time. Many participants used photos to see where to go and took photos to comprehensively view the environment (Lukin et al., 2023).

Participants completed two 20-minute main trials with the robot's assistance after first completing a training trial in which they could familiarize themselves with the interface, activity, and robot capabilities. A paper worksheet was provided to participants, and they were asked to count and photograph objects of interest which varied by trial (e.g., cones, shovels, and shoes). Furthermore, participants were asked to answer the following assessments: "Is there anything that indicates the environment has recently been occupied?", "Were the last occupants speakers of English or a foreign language?", and "Is there anything that you could use to coordinate operations or activities in a headquarters type environment?" Participants were not required to photograph supporting evidence encountered for these assessments, but they verbally reported their conclusion to the experimenter at the trial's end, and could cite encountered evidence.

The robot was controlled by Wizards-of-Oz, human confederates standing in for the robot's dialogue and navigation capabilities. The data collected from earlier SCOUT experiments with Wizards was used to incrementally automate the robot's Dialogue Management (DM) in later experiments. The DM, whether it was a DM-Wizard confederate or the automated system, listened to the participant's instructions and either responded to the participant or passed well-formed and executable instructions to a Robot Navigator (RN) Wizard who reported success or problems. The participant was only made aware they were speaking to a 'robot' and not informed of the inner workings of the robot's controls. The dialogue was annotated for Transaction Units (TUs), a dialogue structure annotation demarcating multiple dialogue turns that sequentially contributed to fulfilling the speaker's original intent across all speakers (the participant, DM, and RN) (Traum et al., 2018).

2.2 The Video-SCOUT Dataset

Video-SCOUT consists of 60 robot-centric RGB videos from the main and training trials of SCOUT Experiments 1 and 2. The total video time of the dataset is 20+ hours, averaging 20:14 minutes per video (Table 1). The videos are accompanied by a transcript file containing the dialogue between the participant and robot (DM-Wizard) which can be played as subtitles with the video. Additionally,

	# Videos	Video Length	
		Total	**Average**
Video-SCOUT	*60*	*20 hrs. 13 min.*	*20:14 min.*
Main Trials	40	14 hrs. 36 min.	21:55 min.
Train Trials	20	5 hrs. 37 min.	16:53 min.
TU Videos	1,672	15 hrs. 17 min.	32 sec.

Table 1: Video-SCOUT summary statistics

each video is split into its constituent TUs, totaling 1,672 TU video clips across main and training trials.

The *robot videos* were created by first extracting the `sensor_msgs/rgb/image_raw` topic from the SCOUT experimental bag files, then using `ffmpeg`[1] to create `.mp4` files. The average length of the main trial videos is 21:55 minutes (max: 27:28 due to network technical difficulties, min: 20:41) and the training trials average 16:53 minutes (max: 22:37, min: 9:56). All videos include an initial calibration procedure and the participant verbally reporting their counting and assessment responses to the experimenter at the conclusion of each trial. The robot had access to these videos during the experiment, while the participant could only see the robot's in-the-moment view with a photo request.

To create *subtitle files*, the time-aligned transcripts from the SCOUT dataset were converted into `.srt` text files with the participant and DM-Wizard dialogue. Each utterance was assigned a sequential ID in the `.srt` file. The participant's push-to-talk keypress timestamps determined how long the utterance subtitle remained on screen. The DM-Wizard text messages remained on screen for 5 seconds after the message was sent to the participant. Subtitle files can be added as a subtitle track to the videos using media players, revealing the robot's view of the environment in real time alongside the ongoing dialogue. An excerpt of the subtitle file from p2.02's main2 trial is below, where 'CMD' ('Commander') indicates the participant:

```
51
00:07:14.48 --> 00:07:19.48
CMD: "can you move several yards towards the
white door"
52
00:07:22.11 --> 00:07:27.11
Robot: "processing. . ."
53
00:07:53.12 --> 00:07:58.12
Robot: "I will move forward 6 feet, ok?"
54
00:07:59.63 --> 00:08:04.63
CMD: "uh i think six feet is too fff far"
55
00:08:04.46 --> 00:08:09.46
CMD: "maybe three feet"
56
00:08:08.45 --> 00:08:13.45
Robot: "ok"
57
00:08:30.54 --> 00:08:35.54
Robot: "moving. . ."
58
00:08:36.93 --> 00:08:41.93
Robot: "done"
59
00:08:38.42 --> 00:08:43.42
CMD: "can you take a photo"
60
00:08:43.33 --> 00:08:48.33
Robot: "sent"
```

Transaction Unit (TU) video clips were created by segmenting the trial video into clips containing the beginning and end of each TU. TU annotations were obtained from the SCOUT dialogue structure `.xlsx` spreadsheets.[2] The average length of the TU video clips is 32 seconds (max: 4 min. 9 sec., min: 4 sec.). Accompanying each TU video is a *TU video clip subtitle file* containing the dialogue exclusive to that TU, supplying the robot's view when participant instructions are issued, and revealing successes and discrepancies in common ground when the participant did not have full access to the video like the robot did. The timestamp for the TU's first utterance was set to 0:00:00 to play with the TU video, and the `.srt` utterance ID restarted within each TU. The excerpt above contains two TUs, therefore has two separate TU videos and transcripts (see Appendix A for the `.srt`s.)

3 Comparison with Related Work

Video understanding is typically conditioned on the categorization or detection of an 'event' visible within the video. There are differing levels of granularity in defining an event. At a high level, videos have been assigned a label depicting an overall category of the content, such as 'vehicle' or 'nature' (Thomee et al., 2016), or 'news' or 'travel' (Abu-El-Haija et al., 2016). At a more detailed level, videos have been labeled by their activity in full, such as 'changing a vehicle tire' (Smeaton et al., 2006) or 'poking a hole into a substance' (Goyal

[1] https://ffmpeg.org/

[2] TUs solely between the participant and experimenter are excluded.

et al., 2017). Others take a hierarchical approach, constructing an event template with sub-events and entity roles. A cooking video is labeled with sub-events including 'grill the tomatoes' followed by 'add oil to a pan' (Zhou et al., 2018). Disaster videos have been annotated to identify the who, what, when and where of their sub-events, such as differentiating within the video between emergency responders and people affected by a flood (Sanders et al., 2024). Many of these datasets are accompanied by natural language, including news articles written about the videos, creating rich, multimodal and multilingual datasets for event understanding, e.g., Sanders and Van Durme (2024); Kriz et al. (2025). Commonly, these event-centric tasks examine videos from YouTube, Flickr, Vimeo, movies, and TV shows. The reader can refer to Sanders and Van Durme (2024) for a comprehensive overview of recent event-centric video datasets. Kriz et al. (2025, Table 1, p2) report these videos range in length from 4 seconds to 8 minutes.

The Video-SCOUT dataset of robot-recorded videos has unique content and characteristics compared to these event-centric datasets. 'Events' at any level of granularity do not appear in Video-SCOUT, as there is no human activity depicted nor is there motion within the environment. Furthermore, the quality of Video-SCOUT videos is degraded by low-quality recording devices, and the robot's low-to-the-ground video perspective differs from human height. The average length of a Video-SCOUT video is 20 minutes, challenging how well video understanding can be conducted over a long period of time as opposed to shorter clips and in significantly different domains.

Other video understanding approaches are pattern-based, looking not to apply an 'event' annotation to a video, but rather identify where in the video a pre-defined pattern breaks. These video datasets are based around observing crowded scenes from a stationary camera over a period of time, e.g., the ShanghaiTech Campus Dataset (Luo et al., 2017), the UCSD Pedestrian Dataset (Li et al., 2013), the Subway Dataset (Adam et al., 2008), and the CUHK Avenue Dataset (Lu et al., 2013). They seek to identify anomalies, when the pattern changes, such as people fighting (Adam et al., 2008), or non-pedestrian entities entering a walkway (Pinggera et al., 2016). These videos are similar to Video-SCOUT in that our videos are recorded from a non-human perspective, yet the key difference concerns spatial movement. In Video-SCOUT, it is the robot moving, rather than entities within the environment, thus, as is the case in prior works, we must again look for how to define pattern breaking anomalies within this domain.

In an attempt to generalize an 'event,' other video understanding approaches have instead created a highlight reel (i.e., a set of individual frames extracted from a video compiled into a short video clip) or a video summary tailored for *any* video content (Song et al., 2015; Sul et al., 2023; Chang et al., 2025). Highlight detection algorithms use accompanying language from the video's metadata to extract frames relevant to the text, for example, the YouTube video's title "Killer Bees Hurt 1000-lb Hog in Bisbee AZ" (Song et al., 2015, p1) is used to condition video frame extraction relating to these keywords. To address the lack of typical 'events' in Video-SCOUT, we leverage Chang et al. (2025)'s Aha! highlight detection model and explore how other language inputs may guide video analysis of robot-recorded videos. Chang et al. (2025)'s paper analyzed eight minutes of a robot-recorded video with the input 'what objects are here?' The preliminary analysis suggested Aha! may be used on out-of-domain videos without fine-tuning, yet the paper did not conduct a thorough evaluation.

Video-SCOUT captures explorations of a space that an embodied agent, i.e., a robot, moves through. These embodied explorations are common in human-agent or human-robot exercises taking place in the physical or virtual world, where the agents are given a directive, e.g., move to a specific object or location, that the agent will complete (Das et al., 2018; Shridhar et al., 2020; Majumdar et al., 2024; Bowser et al., 2025). Dialogue within these environments allows the robot to request clarification of ambiguous instructions or referents to resolve ambiguity (Gervits et al., 2021). However, due to the fact that many real-world remote exploration contexts are bandwidth-limited, full information about the environment (including real-time streaming video) may not be available to the human issuing the robot its navigational instructions. The human must make decisions with potentially incomplete information, decisions that the robot must carry out or clarify. We formulate a new video understanding challenge around the uniqueness of our domain: in "understanding" a video without events or pattern breaking anomalies in the human-robot collaborative context.

4 Non-event Oriented Video Assessments

The Video-SCOUT dataset presents new opportunities for advancing research in human-robot collaboration by leveraging out-of-domain video datasets and task requirements. Given that the videos are of real-world quality (i.e., occasionally dark or blurry), and that they are not streamed live to the human, it is possible the human may overlook or miss something important over the course of their exploration. This is a critical opportunity for the robot teammate to provide support by analyzing its constantly running RGB camera. A robot may additionally augment a human's understanding of the long-form retrospective exploration videos in which there are periods of time where nothing is "happening." However, as previously discussed, the Video-SCOUT environment contains few action events. The robot's journey only shows still objects, and thus, new criteria must be defined.

We propose a new challenge task in video understanding on long-form videos which lack canonical 'events.' By reframing what it means to identify an 'event,' we instead analyze the environment depicted within, according to a high-level assessment question. **Non-event Oriented Video Assessments (NOVA)** is designed as a video frame retrieval experiment in which frames highly relevant to NOVA-questions are selected by an algorithm. We measure the success of this task with precision—the number of selected frames relevant to the assessment—as well as recall—the amount of relevant evidence selected from the environment of all possible relevant evidence. Video-SCOUT supplies videos for the NOVA-questions assigned to the participants: "Is there anything that indicates the environment has recently been occupied?", "Were the last occupants speakers of English or a foreign language?", and "Is there anything that you could use to coordinate operations or activities in a headquarters type environment?"

4.1 Approach

We apply two approaches for video frame retrieval tailored to NOVA-questions. The first is Aha![3], the online highlight detection algorithm from Chang et al. (2025) which achieved 91.6% in top-5 mAP (mean Average Precision) with no fine-tuning on large-scale YouTube datasets, outperforming other approaches at its time of writing. There is no formal definition of an event or activity that the algorithm looks for, providing a solid candidate for our assessment task. The algorithm assigns a relevance score to each frame with respect to the video's title, and returns a list of frames above a certain threshold. Aha! operates sequentially, not needing access to future frames to make its determination. These frames together constitute Aha!'s *Selected Frame Album* for a particular video given a NOVA-question. We ran Aha! using light modifications of the NOVA-questions as the natural language input (exact wording in Appendix B) on an NVIDIA RTX 4090.

The second approach is a general VLM, Google Gemini 2.5 Pro[4]. We give as input the full-length video, and prompt it to list timestamps of frames supporting the NOVA-question (prompt in Appendix B).[5] The authors of this paper then extracted the frames from the provided timestamps to create Gemini's *Selected Frame Albums*. Figure 1 shows a subset of selected frames, with complete albums in Appendix C.

4.2 Annotation

Annotation was conducted by the authors of this paper after familiarization with the SCOUT environment and NOVA-questions. To begin, we created an inventory of every object and area within the experiment environment that could reasonably be used to support each NOVA-question, e.g., the table with office chair and newspapers in Figure 1a might support the assessment that the space was used as a headquarters. The total inventory of these objects and areas is listed in Appendix D. As the NOVA-questions are somewhat subjective, annotation was more lenient and inclusive of different possibilities. When participants answered the questions after their trial, they were not scored on if they accounted for all possible evidence, only that they came to a conclusion. For example, when answering whether the participant believed the environment had been recently occupied, one participant answered, "Yes. There is a box of cereal in one room." Our inventory counted the cereal box and several other objects as appropriate evidence.

[3]Model accessed from publicly available repository at https://github.com/aiden200/Aha-/tree/rebuttal and system defaults accepted.

[4]Model accessed between February–March 2026 using Ask Sage and the Ask Sage Persona with temp=0.

[5]At the time of writing, VLMs are unable to process videos in an online manner, unlike Aha!. We discuss the implications of implementation further in Section 4.4, and focus here on testing the relevant frame identification capabilities.

(a) *Relevant* frame for Occupied, Language, and Headquarters (table, chair, cup, newspapers, bottle).

(b) *Relevant* frame for Occupied (plant). *Distractor* frame for Language and Headquarters.

(c) *Distractor* frame for Occupied, Headquarters, and Language.

Figure 1: Selected frame examples and annotations for NOVA-Questions

Instead of our NOVA algorithms providing the *conclusion* in a natural language statement as the participants did, we designed our experiment to exhaustively analyze the environment for *any* possible supporting evidence, taking the form of a set of selected frames. Annotation of the *Selected Frame Albums* was conducted by one annotator (an author of the paper) after the inventory was finalized through group discussion, and was verified by another author. Each frame in Aha! and Gemini's *Selected Frame Albums* were assigned a label based on the NOVA-question: *Relevant* if at least one piece of evidence in that frame was present (e.g., Figure 1a is *Relevant* for all NOVA-questions), or *Distractor* if no evidence was present in that frame (e.g., Figure 1b is a *Distractor* for the Language and Headquarters NOVA-questions, and Figure 1c is a *Distractor* for all NOVA-questions).

To determine the added value each algorithm provides to a human's analysis, it is necessary to measure what the human reasonably could have deduced on their own. To calculate this, the visual content of the participant's photo requests was examined. This represented the total possible knowledge of the environment the participant could have obtained by the end of their trial, because they did not have access to the live video stream. A *Relevant* or *Distractor* label was assigned to the photos regarding each NOVA-question.

Within each *Relevant* frame in the *Selected Frame Albums* and in the participant's photo requests, the annotator indicated which items or areas provided evidence for the NOVA-question from the assessment inventory. Finally, the annotator reviewed each full-length robot video to count the maximum number of observed evidence within that trial. This was conducted to avoid penalizing recall if the robot never passed by particular evidence cues. In this way, if a participant never instructed the robot to explore the room with the table with office chairs, the *Selected Frame Albums* and photo request annotations would not be penalized for not having seen this space. The maximum evidence score was used to compute a customized recall of the *Selected Frame Albums* and the participant's requested photos.

Cohen's Kappa was computed for the *Relevant* and *Distractor* frame annotation within the Participant Photo Requests. Across all NOVA-questions, $\kappa = 0.74$. Agreement varied by NOVA-question (Occupied $\kappa = 0.97$; Language $\kappa = 0.62$; Headquarters $\kappa = 0.49$). Adjudication revealed discrepancies in overlooking items that were mutually agreed to be relevant, e.g., wall signs appearing in multiple frames. Additionally, Cohen's Kappa was computed for the object inventory within each *Relevant* frame from the participant's photo requests. Across all NOVA-questions, $\kappa = 0.85$. Again, agreement varied by NOVA-question (Occupied $\kappa = 0.91$; Language $\kappa = 0.61$; Headquarters $\kappa = 0.92$). Adjudication revealed challenges pertaining to the Language task in accounting for illegibility due to photo noise, distance from the camera, dark photos, and bad angles, i.e., edge-on shots. A consensus was reached regarding edge-on shots: regardless of how powerful a computer vision algorithm or human eyesight, the text cannot be read if it is not properly in the frame, therefore frames with, for example, the calendar shown from the side perspective, were not counted. Regarding the required quality of the text for legibility, annotation was based on human-determination rather than how good the algorithms may be at OCR (Optical Character Recognition). Because the models were only returning the frame, it is up to the human (the annotators or a future user) to determine its useful-

ness. Furthermore, the Language task only asked to determine the language, not fully read the text; therefore, if an annotator determined enough letters or characters were legible, the text was counted. We discuss this more in the Limitations section, and frame annotations with visuals are given in Appendix C. Other discrepancies in annotation were agreed to be oversights, and annotation was revised on both *Relevant* frames and object inventory until agreement was reached. Subsequently, with the new adjudicated guidelines, annotation was verified and adjusted on Aha! and Gemini's *Selected Frame Albums*.

4.3 Results

We examined the 20 main trials from SCOUT's Experiment 1 (ten participants completed two main trials each). A *Selected Frame Album* was created for each NOVA-question (Occupied, Language, Headquarters) for each approach (Aha!, Gemini). This resulted in a total of 60 *Selected Frame Albums*.

	Part. Photos	**Aha!**	**Gemini**
% Relevant Frames	57.23	63.43	90.21
Occupied	79.27	85.53	100
Language	39.70	40.87	89.86
Headquarters	52.73	63.89	80.77
% Distractor Frames	42.77	36.57	9.79
# Photos Requested or Album Length	33.9	12.73	6.03

Table 2: Average precision of participant photos and *Selected Frame Albums*

Table 2 shows the average percent of *Relevant* and *Distractor* annotations. Of participant photos (first column), 57.23% were relevant, with scores varying based on the NOVA-question. These scores match with observations from prior work in which participants used photo requests to answer NOVA-questions and for navigation (refer to Section 2.1 and Lukin et al. (2023) for photo strategies). The other columns show Aha! and Gemini's *Selected Frame Albums* scores. Across all NOVA-questions, Gemini showed high precision: 90.21% of the frames selected by Gemini were annotated as containing relevant evidence, whereas Aha! averaged 63.43% precision. Within NOVA-questions, both Aha! and Gemini saw drastic differences. For Aha!, the Language NOVA-question had the lowest precision score (40.87%), compared to the Occupied and Headquarters NOVA-questions (85.53% and 63.89% respectively). Meanwhile, Gemini achieved 100% precision on the Occupied NOVA-question, compared to 80.77% on the Headquarters NOVA-question and 89.86% on the Language NOVA-question. The two approaches differed further in the length of the *Selected Frame Albums*, with Aha! selecting an average of 12.73 frames from the full-length video, and Gemini an average of 6.03 frames. On average, 33.9 photos were taken by participants.

Table 3 is organized around the individual markers of evidence, measuring recall conditioned on the total number of evidence items that could possibly be found within the full-length video. The first column in Table 3 shows the percentage of evidence found in the participant photo requests. Across all NOVA-questions, the average recall is high at 87.31%, representing the highest possible score the participant could have achieved without having access to the full robot video. Columns *Aha!* and *Gemini* report their respective *Selected Frame Albums* relevance recall. On their own, recall is drastically lower than the participant photos, with Aha! achieving an average recall of 59.50%, and Gemini 54.42%. We interpret the participant's high recall as a result of taking photos to aid navigation, whereas Aha! and Gemini were tasked only to complete the assessments. The participant photos exhibit a trade-off in high recall at the cost of lower precision, whereas Aha! and Gemini are more balanced.

The columns *Part. Photos + Aha!* and *Part. Photos + Gemini* in Table 3 report how much of a recall boost could have been achieved if the participant's photo requests were combined with the automatically created *Selected Frame Albums*. These were computed by taking the unique complement of the object and area inventory between the participant's photo requests and the approaches' *Selected Frame Albums*. This evidence complement yields gains of 2.39% for Aha! and 2.35% for Gemini averaged across NOVA-questions, showing that these *Selected Frame Albums* contribute a small set of unique evidence not observed by the participant in their exploration. As an example of this, in one trial, a participant instructed the robot to move down a hallway, and a sleeping bag and a movie poster were passed in the process. The participant did not photograph these mid-movement views, but these frames were selected from the full video by Aha! and by Gemini respectively.

% Relevance Recall	Part. Photos	Aha!	Gemini	Part. Photos + Aha!	Part. Photos + Gemini
All Assessments	**87.31**	**59.50**	**54.42**	**89.70 (+2.39)**	**89.66 (+2.35)**
Occupied Assessment	89.37	67.62	57.39	92.71 (+3.34)	89.93 (+0.56)
Language Assessment	79.17	52.71	56.16	82.24 (+3.07)	84.89 (+5.72)
Headquarters Assessment	93.39	58.18	49.71	94.16 (+0.77)	94.16 (+0.77)

Table 3: Evidence recall from the participant's photos and Aha! and Gemini's *Selected Frame Albums* (first three columns). The new recall percentage achieved by combining the unique instances of evidence found in participant's photos and the *Selected Frame Albums*, with gains in parentheses (fourth and fifth columns.)

4.4 Discussion

Model Performance. A precision-recall trade-off manifested in our evaluation. Of Gemini's *Selected Frame Albums*, 90.21% of frames were relevant to the NOVA-question. However, and possibly because it only extracted an average of 6.03 frames, its ability to capture all possible evidence in the video was lower, only 54.42%. The metrics leaned differently for Aha! It achieved a lower average precision of 63.43% and recall of 59.50%. The Language NOVA-question was particularly challenging for Aha! (40.87% precision), whereas Gemini achieved 89.86% precision, showing significant ability to identify texts in the videos. On the other hand, Aha! achieved almost 5% higher recall than Gemini on average, and in particular, about 10% recall increase on the Occupied and Headquarters NOVA-questions. Despite these trade-offs in recall and precision, the models' overall gains in supplemental evidence of participant photos is comparable: 2.39% for Aha! and 2.35% for Gemini. This suggests there may be different opportunities for employing Aha! or Gemini for the NOVA task, given that they can perform well under different conditions for different NOVA-questions.

We outline several strategies to increase and balance precision and recall. Because Gemini's *Selected Frame Album* length was short, future iterations could consider chunking the 20-minute video into smaller segments so that its high precision may still be achieved while increasing recall by combining the results of the chunked videos. Some assessments may be harder for the model to reason about than others; for example, it scored lower in deducing what constituted as a headquarters compared to identifying visible written language.

A key strength of Aha! is its adaptability without fine-tuning. While it achieved a remarkable 91.6% on the event-centric video datasets using the accompanying video titles reported in its paper (Chang et al., 2025), it scored high in *Distractors* within the SCOUT domain (36% of Aha!'s frames were annotated as *Distractor*). To maintain its moderate recall on the Occupied and Headquarters NOVA-questions, we posit that *Distractors* may be minimized with an additional filtering process tailored or fine-tuned to the SCOUT domain. Observationally, most of Aha!'s *Distractors* focused on empty spaces, such as Figure 1c. In both models, the Language NOVA-question scored lower than the others, which, as we discussed in Section 4.2 and more in the Limitations section, comes with legibility challenges.

New HRI Challenges. The identification of frames in non-event videos introduces new opportunities for human-robot collaboration. The high recall scores of both Aha! and Gemini is encouraging and may enable both real-time assisted exploration and retrospective exploration analysis. In the former, it will be critical to minimize disruption and maintain or increase the human's trust in the robot if it is providing real-time alerts on locations possibly relevant to the assessment. A robot with high *Distractors* could quickly erode trust, and after a few incorrectly flagged frames, a human may ignore the robot's suggestions as it interrupts their attempt to complete the assessment. However, with adequate precision and recall checks, a future experiment could allow the robot to take the initiative and offer to stop and look.

There remain gaps in running both Aha! and Gemini in real time. Aha! requires considerable processing power that may be unavailable onboard a robot, and Gemini cannot process videos fully in real time. By contrast, both models would serve well in retrospective exploration analysis. A future experiment could provide a human the *Selected Frame Albums* in addition to their photo requests for greater post-exploration recall. Additionally, it remains to be seen whether a high number of *Distractors* may be less critical in a post-experiment review where real-time interruption is not an issue.

5 Future of Human-Robot Collaboration

NOVA represents a foundational step in supporting human-robot collaboration through video understanding. By enabling a robot to analyze its surroundings with respect to a human's goals, the robot is poised to understand other elements of the collaboration at a more granular level. While the NOVA-question guided the entirety of the *Selected Frame Album* creation, the human is involved in the robot's journey every step of the way by speaking to the robot about what they want to do. There are often cases of a mismatch between the human's language about the environment and what the robot is actually seeing.

We envision future human-robot paradigms which leverage the robot's understanding of a human's instructions in close combination with its visuals to identify when common ground may be lost and to seek strategies to prevent or mitigate it. We thus begin to develop a new challenge leveraging the NOVA framework and the Video-SCOUT dataset called the **Common Ground Alignment Problem (Common-GAP)**. Common-GAP is a decision problem for proactive multimodal repair in bandwidth-limited human-robot exploration. Given the robot's video, the dialogue history, and current dialogue structure annotation, we propose three different challenges for the robot: i) predict misalignment, ii) resolve reference ambiguity, and iii) take initiative. In i), the robot should preemptively detect misalignment in what the human and the robot believe about the world, such as the presence or absence of an object mentioned. In ii) the robot should instigate effective disambiguation and repair strategies to restore common ground. Finally, in iii) the robot should learn when to take initiative and proactively report an observation based on the dialogue history and its continuous video feed. We plan to leverage Video-SCOUT and in particular, the TU Video Clips. See Appendix E for full examples of each decision state.

6 Conclusion

We propose a new challenge in video understanding specific to human-robot collaboration: Non-event Oriented Video Assessment (NOVA). This challenge utilizes our novel Video-SCOUT dataset of 60 robot-recorded, long-form videos with accompanying dialogue transcripts and dialogue structure video clips. Our experiments show promise in using VLMs on out-of-domain videos without fine-tuning, yet reveal gaps in implementation within practical applications. We begin to develop a future task leveraging NOVA and the dialogue structure and video affordances of Video-SCOUT to address the Common-GAP in continuous, human-robot exercises. We invite the community to use this data and contribute algorithms to these challenges.

Limitations

While we designed the Aha! and Gemini prompts to be as similar to the NOVA-Questions as possible, dissimilarities arose in the required input format for each model. Additionally, variants in prompts (e.g., instructions or rephrases of the NOVA-questions) were not exhaustively tested for improved performance. We did not provide examples to the models, operating instead in a zero-shot setting, and thus performance may be improved through future iterations. The full set of inputs and prompts are listed in Appendix B. Furthermore, we did not specify or fine-tune the number of frames for Aha! and Gemini to select. This made the comparison challenging, as Aha! selected approximately twice as many frames as Gemini. We consider what new evaluation metrics are appropriate for this evaluation that are sensitive to the number of frames selected as well as the fact that they are unordered and require different ways to reward prioritization.

Adjudication revealed challenges in assigning annotation in the Language NOVA-question. Aha! and Gemini were only instructed to retrieve the frame or timestamp of a frame to answer the question. We cannot infer understanding of the text visible in the frame on behalf of the model since that was not the question asked of it. In future work, these models may be prompted differently to transcribe the text in images they retrieve to assess model legibility vs. human annotator legibility.

Acknowledgments

This work was supported in part by the U.S. Army Research Laboratory under the Advanced Research Technology, Inc. contract.

References

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. YouTube-8M: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*.

Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz. 2008. Robust real-time unusual event detection using multiple fixed-location monitors. *Transactions on Pattern Analysis and Machine Intelligence*, 30(3):555–560.

Claire Bonial, Stephanie M Lukin, Mitchell Abrams, Anthony Baker, Lucia Donatelli, Ashley Foots, Cory J Hayes, Cassidy Henry, Taylor Hudson, Matthew Marge, Kimberly A. Pollard, Ron Artstein, David Traum, and Clare Voss. 2025. Human–robot dialogue annotation for multi-modal common ground. *Language Resources and Evaluation*, 59(2):1525–1575.

Shawn Bowser, Cynthia Matuszek, and Stephanie Lukin. 2025. Towards integrated multimodal interaction: Merging immersive 3D worlds with language based retrieval for 3D scene understanding. In *Proceedings of the 6th Workshop on Intelligent Cross-Data Analysis and Retrieval*, pages 32–37.

Aiden Chang, Celso de Melo, and Stephanie M. Lukin. 2025. Aha—predicting what matters next: Online highlight detection without looking ahead. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Erin K Chiou, Mustafa Demir, Verica Buchanan, Christopher C Corral, Mica R Endsley, Glenn J Lematta, Nancy J Cooke, and Nathan J McNeese. 2022a. Towards human–robot teaming: Tradeoffs of explanation-based communication strategies in a virtual search and rescue task. *International Journal of Social Robotics*, 14(5):1117–1136.

Manolis Chiou, Georgios-Theofanis Epsimos, Grigoris Nikolaou, Pantelis Pappas, Giannis Petousakis, Stefan Mühl, and Rustam Stolkin. 2022b. Robot-assisted nuclear disaster response: Report and insights from a field exercise. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4545–4552. IEEE.

Hareesh Chitikena, Filippo Sanfilippo, and Shugen Ma. 2023. Robotics in search and rescue (SAR) operations: An ethical and design perspective framework for response phase. *Applied Sciences*, 13(3):1800.

Herbert H. Clark and Catherine R. Marshall. 1981. Definite reference and mutual knowledge. In *Elements of Discourse Understanding*. Cambridge University Press.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10.

Daniel S Drew. 2021. Multi-agent systems for search and rescue applications. *Current Robotics Reports*, 2(2):189–200.

Bernardo Esteves Henriques, Mirko Baglioni, and Anahita Jamshidnejad. 2024. Camera-based mapping in search-and-rescue via flying and ground robot teams. *Machine Vision and Applications*, 35(5):117.

Odair Fernandes, Robin Murphy, David Merrick, Justin Adams, Laura Hart, and Jarrett Broder. 2019. Quantitative data analysis: Small unmanned aerial systems at Hurricane Michael. In *2019 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pages 116–117. IEEE.

Felix Gervits, Gordon Briggs, Antonio Roque, Genki A Kadomatsu, Dean Thurston, Matthias Scheutz, and Matthew Marge. 2021. Decision-theoretic question generation for situated reference resolution: An empirical study and computational model. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 150–158.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. 2017. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5842–5850.

Vimukthi Jayawardene, Thomas J Huggins, Raj Prasanna, and Bapon Fakhruddin. 2021. The role of data and information quality during disaster response decision-making. *Progress in Disaster Science*, 12:100202.

Reno Kriz, Kate Sanders, David Etter, Kenton Murray, Cameron Carpenter, Hannah Recknor, Jimena Guallar-Blasco, Alexander Martin, Eugene Yang, and Benjamin Van Durme. 2025. MultiVENT 2.0: A massive multilingual benchmark for event-centric video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24149–24158.

Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. 2013. Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):18–32.

Cewu Lu, Jianping Shi, and Jiaya Jia. 2013. Abnormal event detection at 150 FPS in Matlab. In *2013 International Conference on Computer Vision (ICCV)*.

Stephanie Lukin, Claire Bonial, Matthew Marge, Taylor A Hudson, Cory Hayes, Kimberly Pollard, Anthony Baker, Ashley N Foots, Ron Artstein, Felix Gervits, Mitchell Abrams, Cassidy Henry, Lucia Donatelli, Anton Leuski, Susan G. Hill, David Traum, and Clare Voss. 2024. SCOUT: A situated and multimodal human-robot dialogue corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14445–14458.

Stephanie M Lukin, Kimberly A Pollard, Claire Bonial, Taylor Hudson, Ron Artstein, Clare Voss, and David Traum. 2023. Navigating to success in multi-modal human-robot collaboration: Analysis and corpus release. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1859–1865. IEEE.

Weixin Luo, Wen Liu, and Shenghua Gao. 2017. A revisit of sparse coding based anomaly detection in stacked RNN framework. In *2017 International Conference on Computer Vision (ICCV)*.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, and 5 others. 2024. OpenEQA: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16488–16498.

Peter Pinggera, Sebastian Ramos, Stefan Gehrig, Uwe Franke, Carsten Rother, and Rudolf Mester. 2016. Lost and found: detecting small road hazards for self-driving vehicles. In *2016 International Conference on Intelligent Robots and Systems (IROS)*.

Kate Sanders, Reno Kriz, David Etter, Hannah Recknor, Alexander Martin, Cameron Carpenter, Jingyang Lin, and Benjamin Van Durme. 2024. Grounding partially-defined events in multimodal data. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15905–15927.

Kate Sanders and Benjamin Van Durme. 2024. A survey of video datasets for grounded event understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7314–7327.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10740–10749.

Alan F Smeaton, Paul Over, and Wessel Kraaij. 2006. Evaluation campaigns and TRECVid. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330.

Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. TVSum: Summarizing web videos using titles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5179–5187.

Jinhwan Sul, Jihoon Han, and Joonseok Lee. 2023. Mr. HiSum: A large-scale dataset for video highlight detection and summarization. In *Advances in Neural Information Processing Systems*, volume 36, pages 40542–40555.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

David Traum, Cassidy Henry, Stephanie Lukin, Ron Artstein, Felix Gervits, Kimberly Pollard, Claire Bonial, Su Lei, Clare Voss, Matthew Marge, Cory J. Hayes, and Susan G. Hill. 2018. Dialogue structure annotation for multi-floor interaction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Gongcheng Wang, Weidong Wang, Pengchao Ding, Yueming Liu, Han Wang, Zhenquan Fan, Hua Bai, Zhu Hongbiao, and Zhijiang Du. 2023. Development of a search and rescue robot system for the underground building environment. *Journal of Field Robotics*, 40(3):655–683.

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

A Transcript files

Below are the TU video clip subtitle files for the TUs in the subtitle excerpt presented in Section 2.2.

TU14:

```
0
00:00:00.00 --> 00:00:05.00
CMD: "can you move several yards towards the
white door"
1
00:00:07.63 --> 00:00:12.63
Robot: "processing. . ."
2
00:00:38.64 --> 00:00:43.64
Robot: "I will move forward 6 feet, ok?"
3
00:00:45.15 --> 00:00:50.15
CMD: "uh i think six feet is too fff far"
4
00:00:49.98 --> 00:00:54.98
CMD: " maybe three feet"
5
00:00:53.97 --> 00:00:58.97
Robot: "ok"
6
00:01:16.06 --> 00:01:21.06
Robot: "moving. . ."
7
00:01:22.45 --> 00:01:27.45
Robot: "done"
```

TU15:

```
0
00:00:00.00 --> 00:00:05.00
CMD: "can you take a photo"
1
00:00:04.91 --> 00:00:09.91
Robot: "sent"
```

B Natural Language Inputs and Prompts

We crafted natural language inputs and prompts to closely preserve the wording given to participants in the SCOUT experiments, which are reprinted below:

- "Is there anything that indicates the environment has recently been occupied?"
- "Were the last occupants speakers of English or a foreign language?"
- "Is there anything that you could use to coordinate operations or activities in a headquarters type environment?"

The Aha! natural language input utilized a randomly selected template from the following list:

- "'[NOVA-Question-Aha]."'
- "What segment of the video addresses the topic '[NOVA-Question-Aha]?"'
- "At what timestamp can I find information about '[NOVA-Question-Aha]' in the video?"
- "Can you highlight the section of the video that pertains to '[NOVA-Question-Aha]?"'
- "Which moments in the video discuss [NOVA-Question-Aha] in detail?"
- "Identify the parts that mention '[NOVA-Question-Aha]."'
- "Where in the video is [NOVA-Question-Aha] demonstrated or explained?"
- "What parts are relevant to the concept of '[NOVA-Question-Aha]?"'
- "Which clips in the video relate to the query '[NOVA-Question-Aha]?"'
- "Can you point out the video segments that cover '[NOVA-Question-Aha]?"'
- "What are the key timestamps in the video for the topic '[NOVA-Question-Aha]?"'

The variable [NOVA-Question-Aha] was abbreviated and reworded from the SCOUT assessments to fit these predetermined templates, and was selected from the following for the appropriate video assessment:

- "the environment has been recently occupied"
- "the written language"
- "evidence of a meeting headquarters"

The Gemini prompt was as follows: "You are an expert video analyst. Review the video and create a list of frames that support the provided hypothesis. Output your results as a list of timestamps. Hypothesis: [NOVA-Question-Gemini]."

The variable [NOVA-Question-Gemini] was selected from the following for the appropriate video assessment:

- "The environment has been recently occupied"
- "The last occupants were speakers of English or a foreign language"
- "There is something which could be used to coordinate operations or activities in a headquarters type environment"

C Selected Frame Albums

Figures 2 and 3 show Gemini and Aha!'s *Selected Frame Albums* for p1.08's main1 trial on the Language NOVA-question.

D Relevant Inventory Lists

Occupied relevant inventory list (27 observations): shoes, cooking items, shopping bag, plants, newspaper, solo cup on chair, conference table, chairs around table, clock, bottle water, calendar, map, monitor, desk, desk chair, sleeping bag, luggage, clothes, cleaning supplies, posters, TV, books by TV, wall signs, construction items, fire extinguisher, water cooler jug, stop sign.

Language relevant inventory list (16 observations; must be legible): cereal box, newspaper, calendar, map, posters, books by TV, room numbers, no smoking sign, plywood writing, cleaning supplies, yellow caution cone, stop sign, blue wall sign, luggage logo, broom logo, orange bucket.

Headquarters relevant inventory list (16 observations): cooking items, newspaper, solo cup on chair, conference table, chairs around table, clock, bottle water, calendar, map, monitor, desk, desk chair, cleaning supplies, wall signs, fire extinguisher, water cooler jug.

E Common-GAP Examples

In Section 5, we proposed a challenge task in which we seek to leverage the robot's understanding of a

(a) Frame at timestamp 00:00. *Distractor* frame (orange bucket not legible).

(b) Frame at timestamp 04:22. *Relevant* frame (yellow caution cone; cleaning supplies not legible).

(c) Frame at timestamp 10:39. *Relevant* frame (cleaning supplies).

(d) Frame at timestamp 13:25. *Relevant* frame (wall sign).

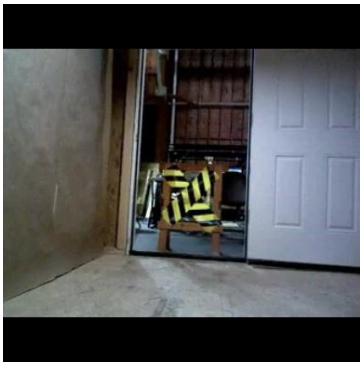
(e) Frame at timestamp 18:21. *Relevant* frame (wall sign).

(f) Frame at timestamp 20:40. *Relevant* frame (blue wall sign).

Figure 2: *Selected Frame Album* assembled by Gemini for p1.08's main1 trial on the Language NOVA-question

(a) Frame at timestamp 02:32. *Distractor* frame (yellow caution cone and orange bucket not legible).

(b) Frame at timestamp 04:18. *Distractor* frame (cleaning supplies not legible).



(c) Frame at timestamp 05:01. *Distractor* frame.

(d) Frame at timestamp 19:51. *Distractor* frame.

(e) Frame at timestamp 20:40. *Relevant* frame (blue wall sign).

Figure 3: *Selected Frame Album* assembled by Aha! for p1.08's main1 trial on the Language NOVA-question

human's instructions in close combination with its visuals to identify when common ground may be lost. We hypothesize that we can use the live robot visual along with the human instruction to automatically detect issues with the instruction. Specifically, we aim to enable i) prediction of misalignment in what the user and the robot believe about the world, ii) resolving reference ambiguity, and iii) taking initiative and proactively reporting an observation. This analysis extends beyond referring expression resolution—the robot must recognize whether or not there is a clear, ambiguous, or even incorrect reference to the environment, and then work within that space or attempt to re-establish common ground. We refer to this challenge to detect and repair as the **Common-GAP: Common Ground Alignment Problem**. Common-GAP is a decision problem for proactive multimodal repair in bandwidth-limited human-robot exploration. Given the current robot video, the dialogue history, and dialogue-structure annotations, decide whether to:

a) execute the command directly,

b) ask a clarification question,

c) offer a photo,

d) send a photo without waiting, or

e) explain why the command cannot yet be grounded.

In an initial exploration of Common-GAP, we manually reviewed the 60 transcripts from SCOUT's Experiments 1 and 2, and selected three reference episodes that parallel the system capabilities described above that we hope to enable. We then viewed the TU video clips alongside the dialogue. We observe cases where the collaborative exploration flows fairly smoothly despite a mismatch in common ground in the form of fully shared visual information. For example, in the following case, the participant would like to enter a room ahead, but there is a road cone blocking the entrance. The offer and subsequently provided visual information make this clear to the participant, who then quickly moves on to explore the other doorways in the environment.

p1.01 training trial, TU 4:

```
CMD: "go in the room"
Robot: "There's an obstruction at the entrance.
  Would you like me to send a picture?"
CMD: "send a picture"
Robot: "sent"
```

p1.01 training trial, TU 5:

```
CMD: "go to the other door"
Robot: "executing..."
Robot: "done"
```

This example showcases how the data might support enabling i): predicting misalignment in what the user and the robot believe about the world, e.g., presence or absence of an object.

We also observe cases where there are longer-horizon miscommunications that persist over several TUs due to the complexity of the physical environment. This is especially prevalent as participants explore areas with multiple adjacent doorways, as is the case in the following example. Again, the supplied visual information enables, over the course of five TUs, the participant to successfully move through the appropriate, disambiguated doorway.

p2.06 main1 trial, TU 4:

```
CMD: "turn around one hundred and eighty degrees
CMD: "and travel through the door"
Robot: "processing..."
Robot: "I will turn around 180 degrees"
Robot: "but..."
Robot: "I'm not sure which doorway you are
  referring to."
```

p2.06 main1 trial, TU 5:

```
Robot: "Should I send a picture?"
CMD: "yes"
Robot: "done, sent"
```

p2.06 main1 trial, TU 6:

```
CMD: "travel straight down the hallway"
Robot: "Which doorway?"
CMD: "four feet"
Robot: "ok. moving..."
Robot: "done"
```

p2.06 main1 trial, TU 7:

```
CMD: "take a picture"
Robot: "sent"
```

p2.06 main1 trial, TU 8:

```
CMD: "travel to the end of the hallway"
CMD: "and enter the doorway on the right"
Robot: "processing..."
Robot: "moving..."
Robot: "done"
```

Thus, this example demonstrates how the data might be used to learn ii): resolving reference ambiguity by instigating effective disambiguation and repair strategies. The above situation and others like it suggest that one strategy may be to automatically provide visual information when there are multiple of the same type of referent in the visual field, such as clustered doorways.

Furthermore, there are cases where automatically sending visual information would be helpful as some operators do not accept offers in dialogue, even when it would be helpful. This is the case in the next interaction, where the participant declines the offer for a picture:

p2.08 training trial, TU 3:

```
CMD: "go ahead.
CMD: "we're looking for doorways"
Robot: "Hmm. . ."
CMD: "go ahead. move for"
```

p2.08 training trial, TU 4:

```
Robot: "Would you like me to send a picture?"
CMD: "no thank you not right now"
```

This user then goes on to struggle with the interaction as exhibited by the fact that the operator issues two more ambiguous, unactionable commands (not shown in the exchange above). This operator only successfully moves to the desired location later, notably after requesting a picture of the environment. Subsequently, the operator leverages a clear pattern of move instructions followed immediately by requests for images, demonstrating the efficacy of this strategy for this particular operator. Thus, this example showcases both the importance and complexity of capability iii): learning when to take initiative and proactively report an observation. From the above example, we see that different kinds of users will react distinctly to different levels of proactive behaviors from the robot—different types of strategies must be leveraged with different operators to supply critical information in a way that maintains conversational norms and expectations.

In developing the Common-GAP and identifying episodic examples from Video-SCOUT, we plan to outline metrics for success, including the time it takes to overcome misunderstanding using the different strategies, to avoid it in the first place, and establishing the threshold for when the robot should employ a strategy. We plan to test different models and ablations (with LLMs, VLMS; with and without domain knowledge, dialogue history, etc.) implementing the Common-GAP on the Video-SCOUT TU clips to detect miscommunication at the earliest point. We will design a human-robot experiment in which such strategies are employed in real time, and human performance and perceptions of the communication are collected.

Less is More: Controlled Visual Evidence Routing and Redundancy Compression for Key Information Extraction

Yang Li[1,3] **Yajiao Wang**[1,3] **Wenhao Hu**[2]
Mengting Zhang[1,3] Zhixiong Zhang[1,3*]

[1]National Science Library, Chinese Academy of Sciences
[2]University of Electronic Science and Technology of China
[3]Department of Information Resources Management,
School of Economics and Management, University of Chinese Academy of Sciences

Abstract

Key Information Extraction (KIE) in visually-rich documents is inherently token-centric, yet prevailing multimodal encoders often fuse dense visual patches with text tokens indiscriminately, which can introduce low-density visual noise, intensify modality competition, and cause robustness collapse under distribution shifts. We propose **OTCR**, a lightweight and architecture-agnostic framework that turns vision from a competitor into a selective supporter for extraction. OTCR learns sparse, interpretable cross-modal coupling via optimal transport to route local visual evidence to the most relevant text tokens, applies token-level gating to control injection strength, and further suppresses spurious correlations through a variational information bottleneck. Experiments on FUNSD, CORD, and SROIE show consistent gains when OTCR is plugged into LayoutLMv3 and GeoLayoutLM, and ablations verify the complementary contributions of coupling, gating, and bottlenecking. Under distribution shifts from Do-GOOD(He et al., 2023a) and EC-FUNSD (Zhang et al., 2024), OTCR markedly mitigates performance degradation, indicating that controlled visual evidence can effectively compensate when text/layout shortcuts become unreliable.

1 Introduction

In visually rich document understanding, key information extraction (Cui et al., 2021) aims to recover structured semantics from unstructured document images, and has demonstrated substantial practical value in scenarios such as invoices, receipts, and reports or forms (Liu et al., 2019; Park et al., 2019; Jaume et al., 2019). Unlike pure text sequence modeling, visually rich documents contain complex two dimensional layouts, diverse typographic styles, and cross modal nonlinear alignment. To capture these heterogeneous signals, Transformer based multimodal pre-trained models (Xu et al., 2020b) introduce two dimensional spatial position

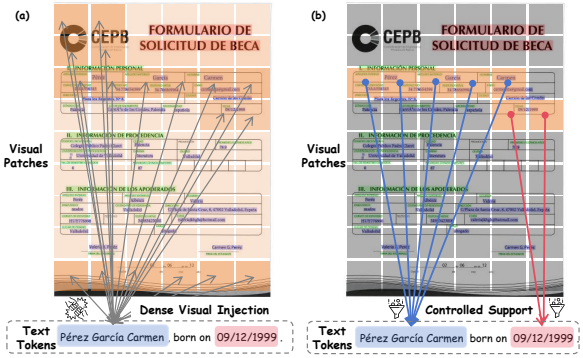


Figure 1: Dense multimodal fusion causes modality competition (a), whereas OTCR performs selective visual support via controlled cross-modal routing (b).

embeddings that explicitly associate text tokens with their bounding box coordinates, enabling the modeling of spatial topology in documents and substantially advancing related tasks. In recent years, research has further improved the upstream pretraining process to learn more general and more unified document representations (Li et al., 2024), thereby supporting a broad range of intelligent visual document tasks, including document classification, layout analysis, and document understanding.

However, what has received limited attention is that, for downstream key information extraction, the task is essentially a fine grained sequence labeling problem that focuses on local regions of a visually rich document. Its prediction ultimately relies on the final layer text representations of the encoder to perform per token classification, where vision serves as supporting evidence rather than the primary semantic carrier (Zhang et al., 2024). As a result, directly introducing page level image patches often incurs a drawback. For field level extraction, the visual evidence truly relevant to a given text token typically comes from a very small local region. In contrast, discretizing the entire page image into a large number of visual tokens inevitably introduces extensive irrelevant background

Proceedings of the 2nd Workshop on
Multimodal Augmented Generation via Multimodal Retrieval, pages 42–53

and spurious cues, including blank areas, textured backgrounds, decorative elements, table rules, and scanning artifacts, causing the visual modality to exhibit a pathological profile of low information density and high noise scale. More critically, visual tokens are dominant in quantity and often highly salient spatially, which can consume the limited attention budget during self attention interactions, thereby triggering modality competition and diluting text to text semantic interactions (Xie et al., 2026).

Furthermore, the aforementioned issues are significantly amplified under distribution shifts and real-world noise scenarios. The Do-GOOD (He et al., 2023a) study indicates that pre-trained visual document understanding models often exhibit a substantial performance gap between in-distribution and out-of-distribution settings. Fine-grained shift analyses reveal the vulnerability of existing models to factors such as template variations, scanning quality degradation, and error accumulation in optical character recognition. The root cause lies in the fact that when templates and layouts change and text or layout shortcuts become unreliable, the model increasingly needs to rely on visual cues for compensation. Concurrently, however, out-of-distribution scenarios introduce a significant increase in blurriness, compression artifacts, background textures, and noise tokens. This makes the visual modality much more susceptible to encompassing pseudo-correlated signals that are irrelevant to the extraction target. Consequently, this phenomenon amplifies the reliance on spurious cues and triggers a collapse in robustness.

To resolve these contradictions, we propose OTCR, a lightweight and architecture-agnostic controlled visual evidence injection framework tailored for Key Information Extraction. It is designed to transform the visual modality from an attention competitor into a selective supporter for text extraction. Our core strategy involves learning sparse and interpretable coupling relationships between text tokens and image patches to achieve structured routing of cross-modal evidence. This mechanism explicitly depicts which visual evidence should serve specific text tokens. Subsequently, we introduce a token-level gating mechanism to dynamically control the intensity of visual injection and suppress the interference of irrelevant visual signals on textual semantic modeling. Finally, we employ a Variational Information Bottleneck to filter and retain complementary information that truly contributes to the task. This enhances robustness without disrupting the text-dominant discriminative structure. Extensive experiments demonstrate that OTCR yields stable gains across multiple mainstream benchmarks and significantly mitigates performance degradation in stress tests involving distribution shifts and layout degradation.

The main contributions of this paper are as follows: (i)We propose OTCR, a lightweight and architecture-agnostic controlled visual evidence injection framework. Starting from the fine-grained nature of the KIE task, this framework successfully reshapes the visual modality from an attention competitor into a selective supporter of text semantics. (ii)We design a multi-level visual control and purification mechanism. By establishing sparse and interpretable cross-modal coupling, we achieve structured routing of visual evidence. Combined with a token-level gating mechanism to dynamically control injection intensity and a Variational Information Bottleneck (VIB) to deeply filter pseudo-correlated noise, we achieve precise retention of complementary information without disrupting the text-dominant structure. (iii)We conduct extensive experimental validation across multiple mainstream KIE benchmarks and backbones. Further out of distribution (OOD) / layout degradation stress tests and case analyses demonstrate that OTCR not only consistently improves task accuracy but also exhibits outstanding generalization capability and stable robustness when confronting complex noise and distribution shift scenarios.

2 Related Works

2.1 Key Information Extraction Method in Visually Rich Documents

Existing KIE methods can roughly be divided into four lines. Early **grid-based approaches** (Katti et al., 2018; Denk and Reisswig, 2019; Kerroumi et al., 2021; Dang et al., 2021) attempted to embed textual semantics directly into a 2D layout space, preserving both content and structure at the input level. LiuGraph (Liu et al., 2019) offered another perspective by modeling documents as **node–edge graphs**, shifting research attention toward more effective graph designs (Biescas et al., 2024; Zhang et al., 2022a; Tang et al., 2021). Subsequently, **large-scale pre-trained models** such as the LayoutLM (Xu et al., 2020b,a; Huang et al., 2022) series, together with more recent **instruction-driven**

MLLM methods (He et al., 2023b; Ye et al., 2023), have unified text, layout, and vision within a single framework, further advancing cross-task generalization. Despite methodological differences, these studies commonly aim to learn a strong multimodal representation to support downstream document intelligence. ViBERTgrid (Lin et al., 2021) integrates BERTgrid (Denk and Reisswig, 2019) with intermediate CNN layers to enable cross-modal interaction; GraphRevisedIE (Cao and Wu, 2023) employs graph revision techniques to combine multimodal embeddings with global contextual information; DocFormer (Appalaraju et al., 2021) leverages carefully designed multi-task unsupervised pre-training to enhance cross-modal alignment; and DocReL (Li et al., 2022) introduces relation consistency modeling to generate more effective relational representations.

2.2 Modality Interference Between Textual and Visual Tokens

Recent VrDU studies have noted that Transformer-style multimodal encoders can suffer from cross-modal interference when heterogeneous token streams, including text, layout, and visual patches, are processed jointly (Nguyen et al., 2021). In such settings, modalities do not contribute symmetrically, and visually salient yet semantically weak regions can disproportionately influence the shared representation space, making token-level linguistic cues that are crucial for extraction-oriented tasks harder to preserve (Zhai et al., 2023). Prior analyses describe this effect through token heterogeneity, visually dominant tokens, and multimodal sequence imbalance, and they consistently characterize the resulting modality competition as a practical bottleneck for stable document IE, particularly in scanned forms and receipts where fine textual distinctions are essential (Zhang et al., 2025).

A complementary line of work studies this issue from an efficiency and robustness perspective (He et al., 2023a). As image resolution increases, the number of visual tokens can grow rapidly, which lengthens multimodal sequences, amplifies interference, and makes cross-modal routing less reliable. Empirical findings further suggest that competition can persist even when visual tokens are not overwhelmingly more numerous, indicating that the issue is not purely a sequence-length effect but also stems from the heterogeneous semantics and salience of multimodal tokens (Toker et al., 2025). Representative directions include sparsifying document structures by pruning graph edges or restricting cross-modal connections, introducing contrastive or consistency objectives to suppress noisy correlations, and designing efficiency-aware tokenization to avoid excessive visual token accumulation (Rombach and Fettke, 2025). While these strategies improve stability in practice, they typically act as implicit regularization rather than providing an explicit mechanism to assign visual evidence to the most relevant textual units and to control how much visual information is retained, leaving room for lightweight and architecture-agnostic frameworks that support controlled evidence transmission and compression for KIE.

3 Proposed Method

3.1 Problem Formulation

Given a visually rich document D, the multimodal inputs consist of a sequence of textual tokens $T = \{t_i\}_{i=1}^{N}$ with corresponding spatial bounding boxes $B = \{b_i\}_{i=1}^{N}$, and a set of discrete visual patches $V = \{v_j\}_{j=1}^{M}$ extracted from the document image. Each bounding box $b_i = (x_i^0, y_i^0, x_i^1, y_i^1)$ provides the 2D geometric coordinates of t_i. The Key Information Extraction (KIE) task is fundamentally formulated as a fine-grained, token-level sequence labeling problem. For each text token t_i, the objective is to predict its corresponding entity label $y_i \in \mathcal{C}$, where $\mathcal{C}$ is a predefined semantic label space. Let $\mathcal{T} \in \mathbb{R}^{N\times d}$ and $\mathcal{V} \in \mathbb{R}^{M\times d}$ denote the initial textual and visual embeddings mapped into a shared feature space.

Standard unconstrained multimodal frameworks directly optimize a dense mapping $\hat{Y} = \Phi(\mathcal{T}, \mathcal{V})$, where textual tokens interact with visual patches in a largely unrestricted manner. However, page-level visual patches often contain low-density and task-irrelevant signals, including blank regions, background textures, decorative elements, table rules, and scanning artifacts. As a result, dense fusion can introduce noisy or spurious visual cues into token representations, increasing modality competition and weakening the fine-grained textual semantics required for token-level KIE.

To formalize a mathematically rigorous filtering process, our OTCR framework defines the extraction pipeline as a constrained latent variable model. Specifically, rather than directly fusing modalities, we aim to learn an intermediate, token-wise latent representation $\mathcal{Z} = \{\mathcal{Z}_1, \ldots, \mathcal{Z}_N\}$ through a con-

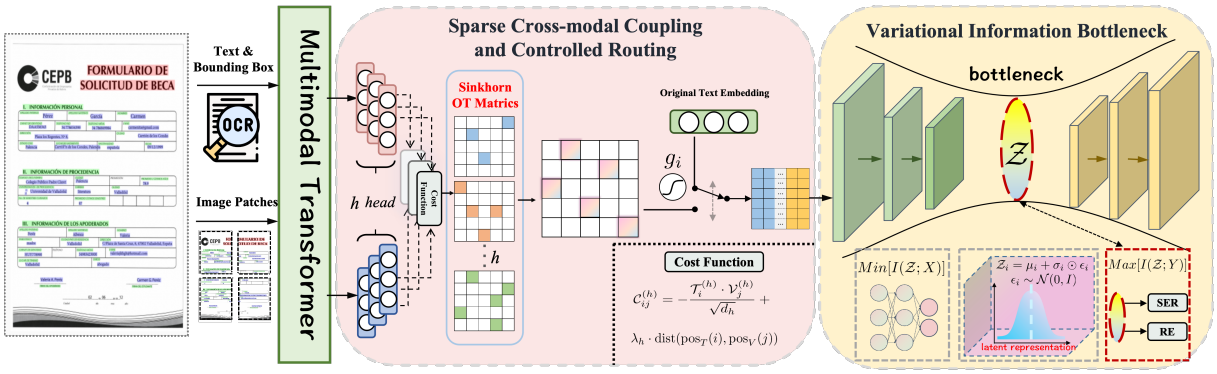


Figure 2: The overall framework of OTCR, which integrates sparse Optimal Transport coupling for controllable visual-to-text injection and a Variational Information Bottleneck for redundancy filtering and task-relevant representation learning.

trolled cross-modal routing function $g(\cdot)$:

$$\mathcal{Z} = g(\mathcal{T}, \mathcal{V}), \quad \hat{y}_i = f(\mathcal{Z}_i; \Theta), \tag{1}$$

where $f(\cdot)$ is the prediction network parameterized by Θ.

3.2 Overview

Our OTCR framework is illustrated in Figure 2. To explicitly prevent visual tokens from cannibalizing the attention budget, Section 3.3 introduces a **Sparse Cross-modal Coupling and Controlled Routing** mechanism. Based on Optimal Transport (OT), this module establishes interpretable alignment paths, guiding visual patches to be selectively injected into textual representations. A dynamic token-level gate is then employed to control the injection intensity, effectively shielding the text semantics from irrelevant visual noise. To further ensure robustness against distribution shifts (OOD), Section 3.4 proposes a **Variational Information Bottleneck (VIB)** training strategy. This module compresses the fused representation, filtering out out-of-distribution pseudo-features (compression artifacts) and retaining only the minimal, task-relevant complementary semantics.

3.3 Sparse Cross-modal Coupling and Controlled Routing

To obtain initial multimodal representations, we first employ an OCR system to extract textual tokens and their 2D bounding boxes. The tokens, layout coordinates, and document image are then fed into a multimodal Transformer, producing text-layout representations $\mathcal{T} \in \mathbb{R}^{N\times d}$ and visual patch representations $\mathcal{V} \in \mathbb{R}^{M\times d}$. Instead of injecting visual patches into text tokens through unconstrained dense fusion, we formulate visual-to-text routing as an entropy-regularized Optimal Transport (OT) problem, which provides a structured and interpretable coupling between textual tokens and visual patches. For the h-th head, we first project the two modalities into a shared subspace:

$$\mathcal{T}^{(h)} = \mathcal{T}\mathbf{W}_h^T, \quad \mathcal{V}^{(h)} = \mathcal{V}\mathbf{W}_h^V. \tag{2}$$

We then define the text-patch matching cost as:

$$\mathcal{C}_{ij}^{(h)} = -\frac{\mathcal{T}_i^{(h)}(\mathcal{V}_j^{(h)})^\top}{\sqrt{d_h}} + \lambda_h \operatorname{dist}\big(\operatorname{pos}_T(i), \operatorname{pos}_V(j)\big). \tag{3}$$

where the first term measures semantic affinity and the second term encourages spatially local coupling.

Given the text and visual marginal distributions $r \in \Delta^N$ and $c \in \Delta^M$, where r_i denotes the transport mass assigned to text token t_i and c_j denotes the transport mass assigned to visual patch v_j, we obtain the OT plan $\pi^{(h)} \in \mathbb{R}_+^{N\times M}$ by Sinkhorn normalization. In our implementation, we use uniform marginals by default, $r_i = 1/N$ and $c_j = 1/M$, so that each text token and visual patch contributes equally before task-driven routing. The transport plan satisfies the marginal constraints

$$\pi^{(h)}\mathbf{1}_M = r, \qquad (\pi^{(h)})^\top \mathbf{1}_N = c. \tag{4}$$

Specifically, the OT plan is computed as

$$\pi^{(h)} = \operatorname{diag}(\mathbf{u}) \exp\left(-\frac{\mathcal{C}^{(h)}}{\tau}\right) \operatorname{diag}(\mathbf{v}), \tag{5}$$

where $\mathbf{u}$ and $\mathbf{v}$ are Sinkhorn scaling vectors chosen to satisfy the above marginal constraints. The resulting plan is a soft coupling matrix rather than a

hard sparse assignment. A smaller τ encourages a more concentrated routing distribution.

We aggregate the plans from all heads and row-normalize them to obtain the final token-wise routing matrix:

$$\mathcal{P} = \mathrm{RowNorm}\left(\frac{1}{H}\sum_{h=1}^{H}\pi^{(h)}\right). \tag{6}$$

The OT-routed visual evidence for token i is computed as:

$$\mathcal{F}_{ot}(i) = \sum_{j=1}^{M}\mathcal{P}_{ij}\mathcal{V}_j. \tag{7}$$

To prevent unreliable visual evidence from overwhelming text semantics, we introduce an entropy-aware token-level gate. We first compute the normalized routing entropy:

$$H_i = -\frac{1}{\log M}\sum_{j=1}^{M}\mathcal{P}_{ij}\log(\mathcal{P}_{ij}+\epsilon). \tag{8}$$

A larger H_i indicates that the visual evidence is more dispersed and therefore less reliable. The gate is then defined as:

$$g_i = \sigma(\mathbf{W}_g[\mathcal{T}_i;\mathcal{F}_{ot}(i)] + b_g - \alpha H_i), \tag{9}$$

where $\alpha \geq 0$ controls the strength of entropy-based suppression.

Finally, the controlled representation is obtained by interpolating between the original text-layout representation and the routed visual evidence:

$$\mathcal{T}_i' = (1-g_i)\mathcal{T}_i + g_i\mathcal{F}_{ot}(i). \tag{10}$$

When the routed visual evidence is concentrated and reliable, the gate allows it to complement the textual representation. When the evidence is scattered or ambiguous, the gate suppresses visual injection and preserves the text-dominant representation.

3.4 Variational Information Bottleneck for Spurious Noise Suppression

While the OT-based gating effectively restricts the intensity of visual injection, the routed representations may still entangle with spurious cues (e.g., scanning artifacts, domain-specific background textures), which are the primary culprits for robustness collapse in OOD scenarios. To guarantee that the final representation is strictly task-oriented, we introduce an intermediate latent representation $\mathcal{Z}$ governed by the Information Bottleneck principle. It aims to be sufficient and minimal: maximizing task-relevant information $I(\mathcal{Z};Y)$ while strictly compressing redundant modality inputs $I(\mathcal{Z};X)$.

Taking the gated representation $\mathcal{T}_i'$ as input, a variational encoder parameterizes a Gaussian distribution for each token:

$$\mu_i = W_\mu\mathcal{T}_i' + b_\mu, \quad \log\sigma_i^2 = W_\sigma\mathcal{T}_i' + b_\sigma. \tag{11}$$

Using the reparameterization trick, we draw the stochastic latent representation:

$$\mathcal{Z}_i = \mu_i + \sigma_i \odot \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0,I). \tag{12}$$

To enforce the bottleneck, the overall optimization process is driven by two components. First, the task supervision loss $\mathcal{L}_{task}$ acts as the variational lower bound to maximize $I(\mathcal{Z};Y)$, formulated as the cross-entropy:

$$\mathcal{L}_{task} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c\in\mathcal{C}} y_{i,c}\,\log\hat{y}_{i,c}. \tag{13}$$

Second, to explicitly formalize the compression of redundancy ($I(\mathcal{Z};X)$), we calculate the Information Bottleneck penalty $\mathcal{L}_{VIB}$ as the Kullback-Leibler (KL) divergence between the posterior distribution and an isotropic Gaussian prior $\mathcal{N}(0,I)$:

$$\mathcal{L}_{VIB} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{KL}\big(q(\mathcal{Z}_i|\mathcal{T}_i')\,\|\,\mathcal{N}(0,I)\big). \tag{14}$$

Finally, the overall optimization objective seamlessly incorporates both terms:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \beta\mathcal{L}_{VIB}, \tag{15}$$

where β is a hyperparameter balancing discriminative power and information compression. See section 4.4.3 for the analysis of β.

By forcing $\mu \to 0$ and $\sigma^2 \to 1$ via $\mathcal{L}_{VIB}$ for task-irrelevant dimensions, the network explicitly penalizes and discards the low-density visual noise and pseudo-features introduced during scanning or template variations. Consequently, the remaining dimensions in $\mathcal{Z}_i$ securely preserve the highly discriminative, text-dominant complementary information, thereby yielding stable KIE performance across both ID and challenging OOD environments.

3.5 Task-Specific Prediction and Inference

After obtaining the purified latent representation $\mathcal{Z}_i$ for each token, we employ lightweight task-specific heads to perform downstream Key Information Extraction, which typically comprises Semantic Entity Recognition (SER) and Relation Extraction (RE).

For the SER task, we apply a multi-layer perceptron (MLP) over the latent representation $\mathcal{Z}_i$ to project it into the predefined label space $\mathcal{C}$, followed by a Softmax activation to obtain the entity probability distribution $\hat{y}_i$:

$$\hat{y}_i = \text{Softmax}(\text{MLP}_{ser}(\mathcal{Z}_i)). \tag{16}$$

For the RE task, which aims to predict the directed linkage between a pair of tokens (key-value pairs), we construct a pairwise representation. Given the latent variables $\mathcal{Z}_i$ and $\mathcal{Z}_j$ of two candidate tokens, we concatenate them and feed them into a relation classifier to predict the linkage probability $\hat{r}_{i,j}$:

$$\hat{r}_{i,j} = \sigma(\text{MLP}_{re}\left([\mathcal{Z}_i; \mathcal{Z}_j]\right)), \tag{17}$$

where σ denotes the sigmoid function. The corresponding relation loss is jointly optimized with the entity classification loss $\mathcal{L}_{task}$ defined in Section 3.4.

Crucially, the behavior of our OTCR framework differs between the training and inference phases. During training, $\mathcal{Z}_i$ is stochastically sampled via the reparameterization trick to enforce the information bottleneck constraint and explore the latent space. However, during the deterministic inference phase, we disable the stochastic noise ϵ and directly utilize the predicted mean μ_i as the definitive latent representation ($\mathcal{Z}_i = \mu_i$). This ensures that the evaluation is stable and strictly relies on the highly selective, text-dominant semantics purified by our controlled routing mechanism.

4 Experiments

4.1 Experiment Settings

4.1.1 Datasets and Baselines

We evaluate OTCR on three standard KIE benchmarks: FUNSD (Jaume et al., 2019), CORD (Park et al., 2019), and SROIE (Huang et al., 2019), covering diverse scanned forms and real-world receipts. To demonstrate the architecture-agnostic nature of our framework, we integrate OTCR into two representative Transformer-based VrDU backbones: **LayoutLMv3** (Huang et al., 2022) (based on ViT-style patch processing) and **GeoLayoutLM** (Luo et al., 2023) (emphasizing geometric pre-training). We fine-tune these models for semantic entity recognition across all datasets and additionally report relation extraction results for FUNSD. Besides the base backbones, we compare OTCR with other representative methods, including graph-based **DocGraphLM** (Wang et al., 2023), structure-aware **LiLT** (Wang et al., 2022a). Following standard protocols, we report the entity-level F1 score for all experiments.

4.1.2 Implementation Details

We employ `LayoutLMv3-base` and `GeoLayoutLM-base` as our primary encoders. Input document images are resized to 224×224 for LayoutLMv3 consistent with its pre-training, while maintaining the original resolution for GeoLayoutLM's geometric extraction. To accommodate document length variations, we set the sequence length to 512 for LayoutLMv3 and 1024 for GeoLayoutLM. For the Sparse Cross-modal Coupling module, the number of optimal transport heads is set to $H = 12/16$, aligned with the backbone's attention heads. We set the entropy regularization coefficient $\tau = 0.1$ for the Sinkhorn algorithm to encourage sparse alignment, and the learnable spatial bias parameter λ is initialized to 1.0. For the Variational Information Bottleneck (VIB), the trade-off hyperparameter β is determined via sensitivity analysis (see Section 4.4.3). All models are optimized using AdamW with a weight decay of 1×10^{-2}. The learning rate is initialized at 2×10^{-5} for LayoutLMv3-based experiments and 1×10^{-5} for GeoLayoutLM-based experiments. We employ a linear warmup for the first 5% of training steps, followed by a linear decay schedule. The batch size is set to 16. Training is conducted for 100 epochs on FUNSD and 50 epochs on CORD and SROIE to ensure full convergence. All experiments are executed on a single NVIDIA A100 (80GB) GPU, with fixed random seeds to ensure reproducibility.

4.2 Main Results

We evaluated OTCR on three KIE benchmarks: FUNSD, CORD, and SROIE, using LayoutLMv3 and GeoLayoutLM as backbones. Table 1 reports the performance of OTCR on three standard benchmarks. The data explicitly shows that integrating

Table 1: Main results on KIE benchmarks (FUNSD, CORD, and SROIE). OTCR is plugged into two representative VrDU backbones, LayoutLMv3 and GeoLayoutLM, and evaluated under the standard SER/RE setting. Best results are marked in **bold** and second-best results are underlined. Results reproduced by us are marked with *.

Type	Method	#Params	FUNSD		CORD	SROIE
			SER	RE	SER	SER
Graph	DocGraphLM (Wang et al., 2023)	base	88.77	-	96.93	-
	GraphDoc (Zhang et al., 2022b)	265M	87.77	-	96.93	98.02*
	mmLayout (Wang et al., 2022b)	large	86.49	-	97.38	97.91
	FormNet (Lee et al., 2022)	large	84.69	-	97.28	-
	DocFormer (Appalaraju et al., 2021)	502M	84.55	-	96.99	-
	MatchVIE (Tang et al., 2021)	base	81.33	-	-	96.57
	RE^2 (Ramu et al., 2024)	base	-	71.76	-	-
	Doc2Graph (Gemelli et al., 2022)	base	-	53.36	-	-
	GraphLayoutLM (Li et al., 2023)	372M	-	-	97.86*	-
Attention	LayoutLMv3 (Huang et al., 2022)	368M	91.81*	79.67*	97.02*	96.10*
	LayoutLMv3 (Huang et al., 2022)	133M	90.85	69.80	95.95*	94.73*
	BROS (Hong et al., 2022)	340M	84.52	77.01	97.28	96.62
	DocTr (Liao et al., 2023)	153M	84.0	73.9	98.2	-
	LiLT (Wang et al., 2022a)	base	88.41	62.76	96.07	-
	LAGaBi (Zhu et al., 2023)	133M	91.00	-	97.05	-
	SERA (Zhang et al., 2021)	base	-	65.96	-	-
	SPADE (Hwang et al., 2021)	base	72.0	41.3	-	-
Pre-trained	GeoLayoutLM (Luo et al., 2023)	399M	91.10	88.06	98.23*	96.93*
	Wukong-Reader (Bai et al., 2023)	470M	**93.62**	-	97.27	**98.15**
	LayoutMask (Tu et al., 2023)	404M	93.20	-	97.19	97.27
	Bi-VLDoc (Luo et al., 2022)	409M	93.44	-	97.84	-
	ERNIE-Layout (Peng et al., 2022)	large	93.12	-	97.21	97.55
	DocReL (Li et al., 2022)	142M	-	46.1	97.0	-
	StrucTexT (Li et al., 2021)	107M	83.09	44.1	-	96.88
Ours	OTCR-LayoutLMv3	133M+30	91.95 (↑1.10)	72.18 (↑2.38)	97.01 (↑1.06)	95.18 (↑0.45)
		368M+30	92.33 (↑0.52)	81.17 (↑1.50)	97.35 (↑0.33)	96.93 (↑0.83)
	OTCR-GeoLayoutLM	399M+30	93.12 (↑2.02)	**88.75** (↑0.69)	**98.63** (↑0.40)	98.08 (↑1.15)

OTCR into both LayoutLMv3 and GeoLayoutLM consistently improves results across all datasets.

On the FUNSD dataset, OTCR-GeoLayoutLM achieved the highest performance in RE 88.75% and competitive SER performance, surpassing other methods, including GeoLayoutLM 91.10% and LAGaBi 91.00%. Similarly, OTCR-LayoutLMv3 improved the SER score to 91.95%, outperforming LayoutLMv3 90.85% and showing the framework's ability to enhance performance even with a smaller model size (133M parameters). For CORD, OTCR-GeoLayoutLM achieved 98.63% in SER, outperforming GeoLayoutLM 98.23% and LayoutLMv3 97.02%. OTCR-LayoutLMv3 (368M) also showed notable improvement, reaching 97.35% in SER, a 0.33% point increase over the baseline. These results indicate that OTCR enhances the ability of the model to extract relevant visual information while minimizing noise, even in complex document layouts such as receipts. In SROIE, OTCR-GeoLayoutLM achieved a SER score of 98.08%, which is very competitive compared to other methods such as LayoutLMv3 96.10% and Wukong-Reader 98.15%. The results confirm OTCR's effectiveness in improving KIE performance, particularly in noisy environments and complex document structures like receipts.

4.3 Ablation Study

The ablation study in Table 2 systematically evaluates the impact of key components in the OTCR framework: Optimal Transport (OT), Gate mechanism, and Variational Information Bottleneck (VIB). For both LayoutLMv3-large and GeoLayoutLM, the results reveal that excluding OT significantly impairs performance, particularly in SER and RE tasks, underscoring the critical role of cross-modal coupling for effective text-visual alignment. The Gate mechanism also contributes notably to performance, especially in reducing visual noise and ensuring relevant visual features are integrated effectively. Its removal leads to further declines, particularly in RE on FUNSD.

While the VIB mechanism provides additional performance benefits by filtering redundant visual features, its absence causes a smaller drop in performance compared to OT and Gate, especially on tasks like SROIE. These findings highlight that OT,

Table 2: Ablation study of OTCR components on two backbones: LayoutLMv3-large and GeoLayoutLM. We report SER on FUNSD/CORD/SROIE and RE on FUNSD. Best results are marked in **bold** and second-best results are underlined.

Backbone	Components			FUNSD		CORD	SROIE
	OT	Gate	VIB	SER	RE	SER	SER
LayoutLMv3-large	–	–	–	91.81	79.67	97.02	96.10
	–	–	✓	91.76	79.69	97.09	96.22
	✓	–	–	91.99	79.94	97.13	96.31
	✓	✓	–	92.21	80.95	97.19	96.55
	✓	–	✓	92.12	80.46	97.10	96.70
	✓	✓	✓	**92.33**	**81.17**	**97.35**	**96.93**
GeoLayoutLM	–	–	–	91.10	88.06	98.23	96.93
	–	–	✓	91.40	88.16	98.22	96.98
	✓	–	–	91.52	88.13	98.35	97.22
	✓	✓	–	92.70	88.69	98.58	97.91
	✓	–	✓	92.43	88.23	98.46	97.63
	✓	✓	✓	**93.12**	**88.75**	**98.63**	**98.08**

Gate, and VIB all play vital roles in enhancing KIE performance, with OT being the most influential in enabling effective cross-modal interaction. The full OTCR model consistently outperforms all ablated versions, demonstrating that the combination of these components is essential for maximizing model accuracy, particularly in complex and noisy document layouts.

4.4 Further Analysis

4.4.1 Robustness Evaluation

In the main experiments, OTCR consistently improves SER and RE on standard KIE benchmarks, showing that controlled visual evidence injection is beneficial under conventional in-distribution evaluation. However, real-world document extraction often involves corrupted text, manual edits, template variation, or different annotation protocols. We therefore conduct two complementary evaluations in Table 3. For distribution-shift robustness, models are fine-tuned on FUNSD and evaluated on FUNSD, OOD_H (human-intervened documents), and OOD_T (text corruption) from Do-GOOD (He et al., 2023a). For annotation-protocol evaluation, we report results under the supervised EC-FUNSD setting (Zhang et al., 2024), which re-annotates FUNSD from an entity-centric perspective.

The results show that OTCR brings clear gains under Do-GOOD shifts, especially when text/layout cues are disrupted. For LayoutLMv3-base, OTCR improves SER F1 by 5.88 points on OOD_H and 3.23 points on OOD_T. For GeoLayoutLM, it improves OOD_H by 5.25 points. These gains

Table 3: **Robustness and entity-centric evaluation.** Models are fine-tuned on FUNSD and evaluated on FUNSD, OOD_H, and OOD_T for distribution-shift evaluation. For EC-FUNSD, models are evaluated under the supervised entity-centric annotation setting. We report SER F1, with values in parentheses denoting absolute gains over the corresponding baseline.

Backbone	Model	FUNSD	OOD_H	OOD_T	EC-FUNSD
LayoutLMv3 (base)	Baseline	90.85	73.25	86.82	82.30
	OTCR	91.95 (+1.10↑)	79.13 (+5.88↑)	90.05 (+3.23↑)	83.56 (+1.26↑)
LayoutLMv3 (large)	Baseline	91.81	80.16	87.95	83.88
	OTCR	92.33 (+0.52↑)	84.33 (+4.17↑)	87.98 (+0.03↑)	83.27 (-0.61↓)
GeoLayoutLM	Baseline	91.10	84.26	89.37	83.62
	OTCR	93.12 (+2.02↑)	89.51 (+5.25↑)	90.14 (+0.77↑)	85.30 (+1.68↑)

support our central claim: selectively routed visual evidence can act as a useful complement when textual or layout shortcuts become unreliable. Under the supervised EC-FUNSD setting, OTCR also improves LayoutLMv3-base and GeoLayoutLM, but slightly decreases LayoutLMv3-large. This suggests that controlled visual injection is generally helpful, but its benefit depends on the backbone and the type of shift. Overall, OTCR provides a lightweight mechanism for making visual evidence supportive rather than competitive in KIE.

4.4.2 Comparison with Large-Parameter Models.

Table 4 compares OTCR with representative large-parameter models, including zero-shot multimodal

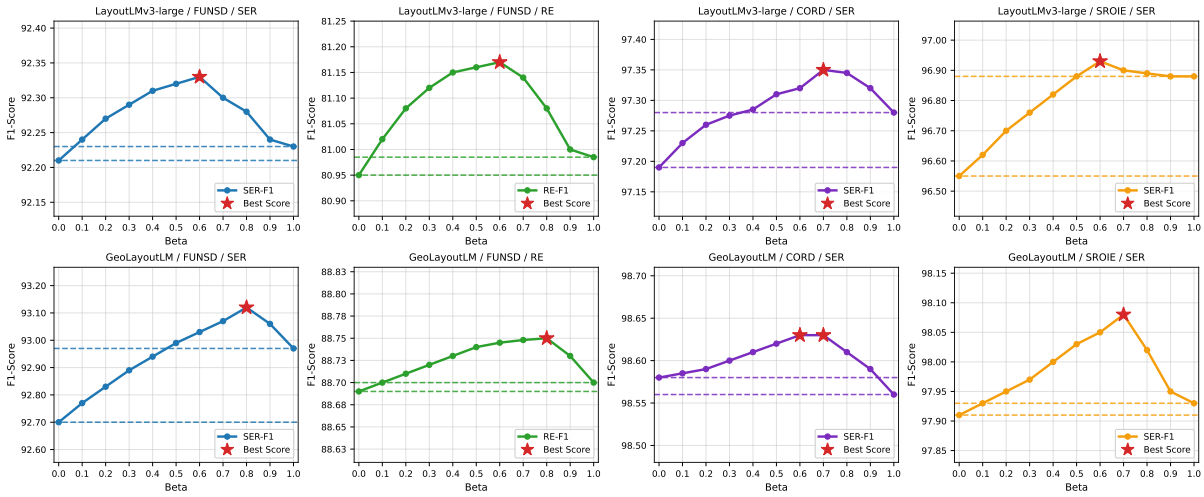


Figure 3: Sensitivity Analysis of β.

Table 4: **Main comparison with large-parameter models.**

Model	FUNSD	CORD	Params	Setting
GPT-5.1	86.78	94.12	–	Zero-shot
Qwen3-VL	79.64	89.93	235B	Zero-shot
Gemini 2.5 Flash	85.52	93.75	–	Zero-shot
LayoutLLM	95.30	98.64	6914.38M	Fine-tune
OTCR-GeoLayoutLM	**93.12**	**98.63**	**399M**+30	**Fine-tune**

LLMs and a finetuned large document model. Despite being built on a lightweight 399M-parameter backbone, **OTCR-GeoLayoutLM** achieves strong finetuned performance, reaching 93.12% on FUNSD and 98.63% on CORD, which is competitive with much larger specialized models (LayoutLLM at 95.30%/98.64% with 6.9B parameters) and substantially exceeds the zero-shot results of MLLMs such as GPT-5.1 (86.78%/94.12%), Gemini 2.5 flash (85.52%/93.75%), and Qwen3-VL (79.64%/89.93%). This highlights that, for extraction-centric VrDU, controlled and task-aligned visual evidence routing can be more effective than scaling alone, enabling small-to-mid scale backbones to approach the performance of multi-billion-parameter systems under supervised finetuning.

4.4.3 Sensitivity Analysis of β

We further analyze the sensitivity of OTCR to the VIB trade-off coefficient β by sweeping $\beta \in [0, 1]$ and reporting the resulting SER/RE curves for LayoutLMv3-large and GeoLayoutLM on FUNSD, CORD, and SROIE (Figure 3). Overall, the performance follows a consistent rise–plateau–slight-drop pattern: increasing β from 0 (no bottleneck) yields steady gains, suggesting that moderate information compression effectively filters redundant multimodal signals and stabilizes token representations, whereas overly large β starts to degrade accuracy due to over-compression. Notably, the optimal β varies slightly across backbones and datasets, reflecting different noise levels and modality redundancy in forms versus receipts; nevertheless, the best-performing region is consistently concentrated around $\beta \in [0.6, 0.8]$ for most settings (peaks marked by $\star$), indicating that OTCR is not overly sensitive to precise tuning and admits a broad, transferable operating range in practice.

5 Conclusion

We propose OTCR, a lightweight and architecture-agnostic framework for controlled visual evidence injection in Key Information Extraction. By introducing sparse optimal transport-based cross-modal routing, token-level gating, and variational information bottleneck compression, OTCR explicitly models modality asymmetry and reshapes vision into a selective supporter of text semantics. Extensive experiments show consistent gains across multiple benchmarks and backbones, with notably improved robustness under distribution shifts.

Limitation

While OTCR demonstrates consistent improvements across multiple benchmarks and backbones, the present study still has several limitations. Our work mainly focuses on extraction-oriented document understanding, and the applicability of con-

trolled visual evidence routing to other visually rich document understanding tasks remains for future investigation. In addition, although OTCR is lightweight and architecture-agnostic, it still introduces several tunable components, and further simplification of the framework may improve its practicality in deployment.

Acknowledgments

The AI-driven experiments, simulations and model training were performed on the robotic AI-Scientist platform of Chinese Academy of Sciences.

References

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 993–1003.

Haoli Bai, Zhiguang Liu, Xiaojun Meng, Li Wentao, Shuang Liu, Yifeng Luo, Nian Xie, Rongfu Zheng, Liangwei Wang, Lu Hou, et al. 2023. Wukong-reader: Multi-modal pre-training for fine-grained visual document understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13386–13401.

Nil Biescas, Carlos Boned, Josep Lladós, and Sanket Biswas. 2024. Geocontrastnet: Contrastive key-value edge learning for language-agnostic document understanding. In *International Conference on Document Analysis and Recognition*, pages 294–310. Springer.

Panfeng Cao and Jian Wu. 2023. Graphrevisedie: Multimodal information extraction with graph-revised network. *Pattern Recognition*, 140:109542.

Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. 2021. Document ai: Benchmarks, models and applications. *arXiv preprint arXiv:2111.08609*.

Tuan Anh Nguyen Dang, Duc Thanh Hoang, Quang Bach Tran, Chih-Wei Pan, and Thanh Dat Nguyen. 2021. End-to-end hierarchical relation extraction for generic form understanding. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5238–5245. IEEE.

Timo I Denk and Christian Reisswig. 2019. Bertgrid: Contextualized embedding for 2d document representation and understanding. *arXiv preprint arXiv:1909.04948*.

Andrea Gemelli, Sanket Biswas, Enrico Civitelli, Josep Lladós, and Simone Marinai. 2022. Doc2graph: a task agnostic document understanding framework based on graph neural networks. *arXiv preprint arXiv:2208.11168*.

Jiabang He, Yi Hu, Lei Wang, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2023a. Do-good: towards distribution shift evaluation for pre-trained visual document understanding models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 569–579.

Jiabang He, Lei Wang, Yi Hu, Ning Liu, Hui Liu, Xing Xu, and Heng Tao Shen. 2023b. Icl-d3ie: In-context learning with diverse demonstrations updating for document information extraction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19485–19494.

Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2022. Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10767–10775.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM international conference on multimedia*, pages 4083–4091.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE.

Wonseok Hwang, Jinyeong Yim, Seunghyun Park, Sohee Yang, and Minjoon Seo. 2021. Spatial dependency parsing for semi-structured document information extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 330–343.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE.

Anoop Raveendra Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2d documents. *arXiv preprint arXiv:1809.08799*.

Mohamed Kerroumi, Othmane Sayem, and Aymen Shabou. 2021. Visualwordgrid: information extraction from scanned documents using a multimodal approach. In *International Conference on Document Analysis and Recognition*, pages 389–402. Springer.

Chen-Yu Lee, Chun-Liang Li, Timothy Dozat, Vincent Perot, Guolong Su, Nan Hua, Joshua Ainslie, Renshen Wang, Yasuhisa Fujii, and Tomas Pfister. 2022.

FormNet: Structural encoding beyond sequential modeling in form document information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3735–3754, Dublin, Ireland. Association for Computational Linguistics.

Qiwei Li, Zuchao Li, Xiantao Cai, Bo Du, and Hai Zhao. 2023. Enhancing visually-rich document understanding via layout structure modeling. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4513–4523.

Qiwei Li, Zuchao Li, Ping Wang, Haojun Ai, and Hai Zhao. 2024. Hypergraph based understanding for document semantic entity recognition. *arXiv preprint arXiv:2407.06904*.

Xin Li, Yan Zheng, Yiqing Hu, Haoyu Cao, Yunfei Wu, Deqiang Jiang, Yinsong Liu, and Bo Ren. 2022. Relational representation learning in visually-rich documents. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4614–4624.

Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. 2021. Structext: Structured text understanding with multi-modal transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1912–1920.

Haofu Liao, Aruni RoyChowdhury, Weijian Li, Ankan Bansal, Yuting Zhang, Zhuowen Tu, Ravi Kumar Satzoda, R. Manmatha, and Vijay Mahadevan. 2023. Doctr: Document transformer for structured information extraction in documents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19584–19594.

Weihong Lin, Qifang Gao, Lei Sun, Zhuoyao Zhong, Kai Hu, Qin Ren, and Qiang Huo. 2021. Vibertgrid: a jointly trained multi-modal 2d document representation for key information extraction from documents. In *International Conference on Document Analysis and Recognition*, pages 548–563. Springer.

Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. 2019. Graph convolution for multimodal information extraction from visually rich documents. *arXiv preprint arXiv:1903.11279*.

Chuwei Luo, Changxu Cheng, Qi Zheng, and Cong Yao. 2023. Geolayoutlm: Geometric pre-training for visual information extraction. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chuwei Luo, Guozhi Tang, Qi Zheng, Cong Yao, Lianwen Jin, Chenliang Li, Yang Xue, and Luo Si. 2022. Bi-vldoc: Bidirectional vision-language modeling for visually-rich document understanding. *arXiv preprint arXiv:2206.13155*.

Laura Nguyen, Thomas Scialom, Jacopo Staiano, and Benjamin Piwowarski. 2021. Skim-attention: Learning to focus via document layout. *arXiv preprint arXiv:2109.01078*.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*.

Qiming Peng, Yinxu Pan, Wenjin Wang, Bin Luo, Zhenyu Zhang, Zhengjie Huang, Teng Hu, Weichong Yin, Yongfeng Chen, Yin Zhang, et al. 2022. Ernie-layout: Layout knowledge enhanced pre-training for visually-rich document understanding. *arXiv preprint arXiv:2210.06155*.

Pritika Ramu, Sijia Wang, Lalla Mouatadid, Joy Rimchala, and Lifu Huang. 2024. Re2: Region-aware relation extraction from visually rich documents. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8723–8739.

Alexander Michael Rombach and Peter Fettke. 2025. Deep learning based key information extraction from business documents: Systematic literature review. *ACM Computing Surveys*, 58(2):1–37.

Guozhi Tang, Lele Xie, Lianwen Jin, Jiapeng Wang, Jingdong Chen, Zhen Xu, Qianying Wang, Yaqiang Wu, and Hui Li. 2021. Matchvie: Exploiting match relevancy between entities for visual information extraction. *arXiv preprint arXiv:2106.12940*.

Michael Toker, Ido Galil, Hadas Orgad, Rinon Gal, Yoad Tewel, Gal Chechik, and Yonatan Belinkov. 2025. Padding tone: A mechanistic analysis of padding tokens in t2i models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7618–7632.

Yi Tu, Ya Guo, Huan Chen, and Jinyang Tang. 2023. Layoutmask: Enhance text-layout interaction in multi-modal pre-training for document understanding. *arXiv preprint arXiv:2305.18721*.

Dongsheng Wang, Zhiqiang Ma, Armineh Nourbakhsh, Kang Gu, and Sameena Shah. 2023. Docgraphlm: documental graph language model for information extraction. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 1944–1948.

Jiapeng Wang, Lianwen Jin, and Kai Ding. 2022a. Lilt: A simple yet effective language-independent layout transformer for structured document understanding. *arXiv preprint arXiv:2202.13669*.

Wenjin Wang, Zhengjie Huang, Bin Luo, Qianglong Chen, Qiming Peng, Yinxu Pan, Weichong Yin, Shikun Feng, Yu Sun, Dianhai Yu, et al. 2022b. mmlayout: Multi-grained multimodal transformer for document understanding. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4877–4886.

Tingwei Xie, Jinxin He, and Yonghong Song. 2026. Roap: A reading-order and attention-prior pipeline for optimizing layout transformers in key information extraction. *arXiv preprint arXiv:2601.05470*.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2020a. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020b. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1192–1200.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. 2023. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*.

Mingliang Zhai, Yulin Li, Xiameng Qin, Chen Yi, Qunyi Xie, Chengquan Zhang, Kun Yao, Yuwei Wu, and Yunde Jia. 2023. Fast-structext: An efficient hourglass transformer with modality-guided dynamic token merge for document understanding. *arXiv preprint arXiv:2305.11392*.

Chong Zhang, Yixi Zhao, Yulu Xie, Chenshu Yuan, Yi Tu, Ya Guo, Mingxu Chai, Ziyu Shen, Yue Zhang, and Qi Zhang. 2024. Unveiling the deficiencies of pre-trained text-and-layout models in real-world visually-rich document information extraction. *arXiv preprint arXiv:2402.02379*.

Jiaxin Zhang, Wentao Yang, Songxuan Lai, Zecheng Xie, and Lianwen Jin. 2025. Dockylin: A large multimodal model for visual document understanding with efficient visual slimming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9923–9932.

Yue Zhang, Zhang Bo, Rui Wang, Junjie Cao, Chen Li, and Zuyi Bao. 2021. Entity relation extraction as dependency parsing in visually rich documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2759–2768.

Zhenrong Zhang, Jiefeng Ma, Jun Du, Licheng Wang, and Jianshu Zhang. 2022a. Multimodal pre-training based on graph attention network for document understanding. *IEEE Transactions on Multimedia*, 25:6743–6755.

Zhenrong Zhang, Jiefeng Ma, Jun Du, Licheng Wang, and Jianshu Zhang. 2022b. Multimodal pre-training based on graph attention network for document understanding. *Trans. Multi.*, 25:6743–6755.

Xi Zhu, Xue Han, Shuyuan Peng, Shuo Lei, Chao Deng, and Junlan Feng. 2023. Beyond layout embedding: Layout attention with gaussian biases for structured document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7773–7784.

KoViDoRe: A Benchmark for Korean Visual Document Retrieval

Yongbin Choi, Yongwoo Song, Mujeen Sung*
Kyung Hee University
{yongbinchoi, syw5141, mujeensung}@khu.ac.kr

Code: https://github.com/whybe-choi/kovidore-benchmark
Dataset: https://hf.co/datasets/NomaDamas/ko-vdr-train-public

Abstract

Recent advances in multimodal retrieval have improved the ability to retrieve information from visually rich documents such as PDFs and reports. However, existing benchmarks remain largely centered on English and provide limited coverage of Korean visual documents with complex structures. Furthermore, most existing Korean resources primarily evaluate single-page retrieval, failing to capture realistic scenarios that require evidence aggregation across multiple pages. To address these gaps, we introduce **KoViDoRe**, a benchmark for Korean visual document retrieval. The dataset is constructed from publicly available Korean documents with diverse layouts, including tables, figures, and multi-column structures. We develop a multi-stage data curation pipeline consisting of structured document parsing, synthetic query generation using both summary-based and context-based strategies, and relevance mapping with human verification. Using KoViDoRe, we evaluate a wide range of multimodal retrieval models and observe that current models struggle to effectively handle Korean visual document retrieval, particularly in settings involving structured content and diverse query types. Motivated by this finding, we further curate a large-scale training dataset, **Ko-VDR Train Public**, to support the development of retrieval models tailored to Korean visual documents. Together, KoViDoRe and Ko-VDR Train Public provide a unified benchmark and training resource for Korean visual document retrieval.

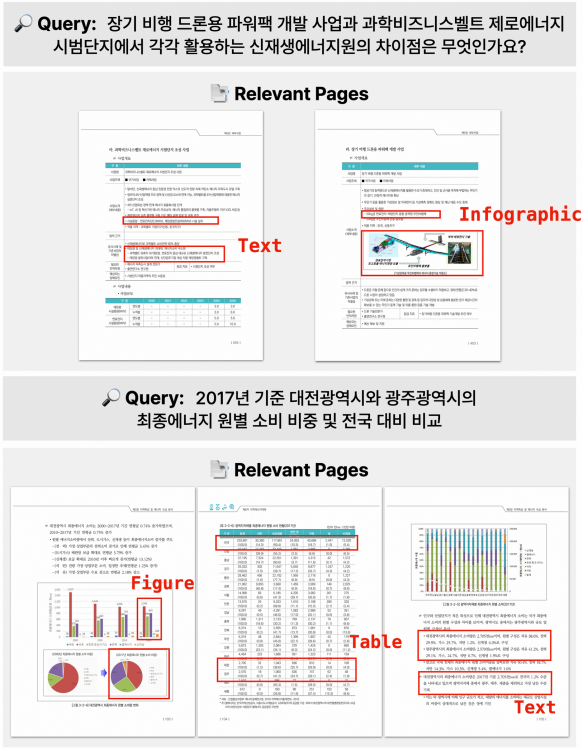


Figure 1: Examples of queries in KoViDoRe and their corresponding relevant pages. Each query requires aggregating evidence from pages and diverse modalities, including text, tables, figures, and infographics.

1 Introduction

Recent advances in multimodal large language models and retrieval-augmented generation (RAG) have significantly improved the ability to retrieve and reason over complex documents (Abootorabi et al., 2025; Song et al., 2025; Yu et al., 2024). In particular, a growing line of work on visual document retrieval (VDR) and multimodal retrieval models, including approaches such as ColPali (Faysse et al., 2024), has demonstrated strong performance in retrieving document pages by jointly modeling textual, visual, and layout information. These developments have enabled systems to move beyond text-only retrieval and better handle structured documents such as PDFs, reports, and forms (Yan et al., 2026). To support this progress, recent benchmarks have adopted large-scale synthetic data generation pipelines, exemplified by frameworks such as ViDoRe (Macé et al., 2025; Loison et al., 2026), enabling scalable evaluation of multimodal retrieval systems. In parallel, efforts such as Jina-VDR (Günther et al., 2025), MIRACL-VISION

*Corresponding author

Proceedings of the 2nd Workshop on Multimodal Augmented Generation via Multimodal Retrieval, pages 54–80

(Osmulski et al., 2025) and SDS KoPub VDR (Lee et al., 2025) have extended this paradigm to non-English settings, providing valuable resources for Korean document retrieval and highlighting the importance of multilingual evaluation.

Despite these advances, existing benchmarks largely formulate VDR as a single-page retrieval task, where each page is treated as an independent unit (Wasserman et al., 2025; Wang et al., 2025). While this formulation simplifies evaluation, it does not accurately reflect how information is organized in real-world documents. In practical scenarios such as financial reports, policy documents, and technical manuals, relevant information is often distributed across multiple pages, requiring systems to aggregate evidence and perform reasoning over a set of pages rather than retrieving a single relevant page (Cho et al., 2024). Although prior datasets may include queries that involve reasoning, such reasoning is typically confined to a single page or limited contextual scopes (Dong et al., 2025). Addressing this limitation requires a shift from single-page retrieval to multi-page evidence aggregation, where retrieval systems must identify a coherent set of pages that collectively fulfill the information need.

In this work, we introduce **KoViDoRe**, a benchmark for Korean visual document retrieval that explicitly models this setting. Building upon prior synthetic data generation approaches, we construct a dataset of realistic enterprise-style documents and generate queries that require retrieving and synthesizing information distributed across multiple pages to provide a complete answer. Unlike existing benchmarks that primarily focus on single-page retrieval with extractive queries, we formulate the task to address multi-page relevance. In this setting, a single query often corresponds to multiple supporting pages, requiring models to identify the full set of pages whose combined content is necessary to satisfy the information need. To enable scalable construction of such queries, we adopt and adapt synthetic query generation techniques to the Korean document domain. Our pipeline leverages large language models to generate queries with diverse reasoning patterns, including multi-hop inference, numerical comparison, and cross-sectional aggregation, while maintaining explicit mappings between queries and their supporting pages. Rather than introducing a fundamentally new generation method, our focus is on restructuring the task and dataset to better reflect realistic retrieval scenarios, particularly in non-English and enterprise contexts.

In addition, we release **Ko-VDR Train Public**, a large-scale training dataset aligned with the proposed task, providing a foundation for developing and evaluating retrieval models in Korean multimodal settings. Through extensive experiments, we show that existing multimodal retrieval approaches struggle significantly under this formulation, especially as the number of required supporting pages increases. These results highlight the limitations of current single-page retrieval paradigms and underscore the need for models that can effectively aggregate over evidence distributed across multiple pages. Our contributions are as follows:

- We introduce **KoViDoRe**, a Korean-focused benchmark designed to evaluate multi-page retrieval performance in realistic document settings.
- We adapt synthetic query generation techniques to construct realistic queries with explicit page-level supervision.
- We release **Ko-VDR Train Public**, a large-scale dataset supporting training in Korean multimodal retrieval.
- We show that existing retrieval models struggle to retrieve evidence distributed across multiple pages on Korean visual documents, highlighting a gap not captured by existing benchmarks.

2 Related Work

2.1 Multimodal Retrieval Models

Recent advances in multimodal retrieval models have significantly improved the ability to retrieve information from visually rich documents (Günther et al., 2025; Ma et al., 2024; Li et al., 2026; Moreira et al., 2026). In particular, models such as ColPali and related late-interaction architectures represent each document page as a unified retrieval unit, encoding the textual content, visual features, and layout structure contained within the page (Faysse et al., 2024; Xiao et al., 2025). This allows retrieval models to capture not only semantic information from text, but also spatial and visual cues, leading to more effective retrieval over visually rich documents. These approaches have demonstrated strong performance across a variety of document understanding tasks, especially in settings where visual structure plays a critical role. In addition to

late-interaction models, dual-encoder and dense retrieval approaches have also been extended to multimodal settings, often leveraging vision-language models to capture both textual and visual semantics (Ma et al., 2024; Nomic Team, 2025). These developments have contributed to substantial progress in retrieving relevant content from structured documents such as PDFs, forms, and reports.

However, despite these modeling advances, the datasets used to train and evaluate such models are largely concentrated on English and European languages (Yu et al., 2024; Günther et al., 2025; Loison et al., 2026; Peng et al., 2025; Wasserman et al., 2025; Shorten et al., 2026). As a result, the performance and behavior of multimodal retrieval models on other languages, including Korean, remain underexplored. This is particularly important in document retrieval settings, where linguistic characteristics such as morphology, spacing variation, and domain-specific expressions interact with visual structure and layout. The lack of dedicated Korean benchmarks limits the ability to assess and develop retrieval models for realistic Korean document scenarios.

2.2 Vision Document Retrieval Benchmarks

Recent benchmarks for visual document retrieval and document-centric multimodal retrieval, such as ViDoRe (Macé et al., 2025; Loison et al., 2026), Jina-VDR (Günther et al., 2025), REAL-MM-RAG (Wasserman et al., 2025), UniDoc-Bench (Peng et al., 2025), MIRACL-VISION (Osmulski et al., 2025), and IRPAPERS (Shorten et al., 2026) have significantly expanded evaluation settings by incorporating visually rich documents, multimodal signals, and realistic query formulations. Many of these benchmarks also support multilingual evaluation. However, their language coverage is largely centered on English and European languages, leaving Korean relatively underrepresented despite its distinct linguistic and document characteristics.

Jina-VDR, for example, includes a Korean subset and broadens the diversity of visual documents and query types. However, its document collections are constructed to cover a wide range of modalities and scenarios, which can make them less representative of real-world Korean document retrieval settings, such as structured public documents, reports, or enterprise-style materials. Similarly, MIRACL-VISION provides multilingual evaluation with Korean queries and documents, but its corpus is primarily derived from Wikipedia, which differs substantially from the types of structured and visually complex documents commonly encountered in real-world Korean retrieval scenarios. This discrepancy limits its ability to fully capture the challenges of practical document retrieval in Korean contexts. SDS KoPub VDR (Lee et al., 2025) addresses this gap by introducing a large-scale benchmark for Korean visual document retrieval. It provides an important step toward evaluating retrieval models on Korean documents with complex layouts and diverse structures. Nevertheless, its query formulation remains strictly focused on single-page retrieval, where each query is mapped to only one relevant page, rather than addressing information distributed across multiple pages.

In contrast, KoViDoRe is designed to emphasize more complex information needs that cannot be satisfied by a single page alone. Our queries require aggregating evidence across multiple pages and capturing relationships between distributed pieces of information. By focusing on Korean documents while introducing queries with higher reasoning complexity and more realistic document distributions, KoViDoRe complements existing benchmarks and provides a more challenging evaluation setting for Korean visual document retrieval.

3 Dataset Curation

As illustrated in Figure 2, our dataset curation pipeline consists of document collection, structured parsing, query generation, and relevance mapping. The overall pipeline design is inspired by ViDoRe V3 (Loison et al., 2026), and is adapted to better reflect the characteristics of Korean document ecosystems and real-world data sources.

3.1 Source Collection

To construct a realistic benchmark for Korean visual document retrieval, we collect document corpora from publicly available sources, including government reports, policy documents, and enterprise-style materials obtained from the Korean public data portal[1] and official institutional websites. We prioritize documents that are freely available under the Korea Open Government License (KOGL) Type 1 and 2, as well as materials without restrictive licensing conditions. This ensures that the collected data can be used for research and downstream applications without legal constraints. Following SDS KoPub VDR (Lee et al., 2025), which

[1] https://www.data.go.kr

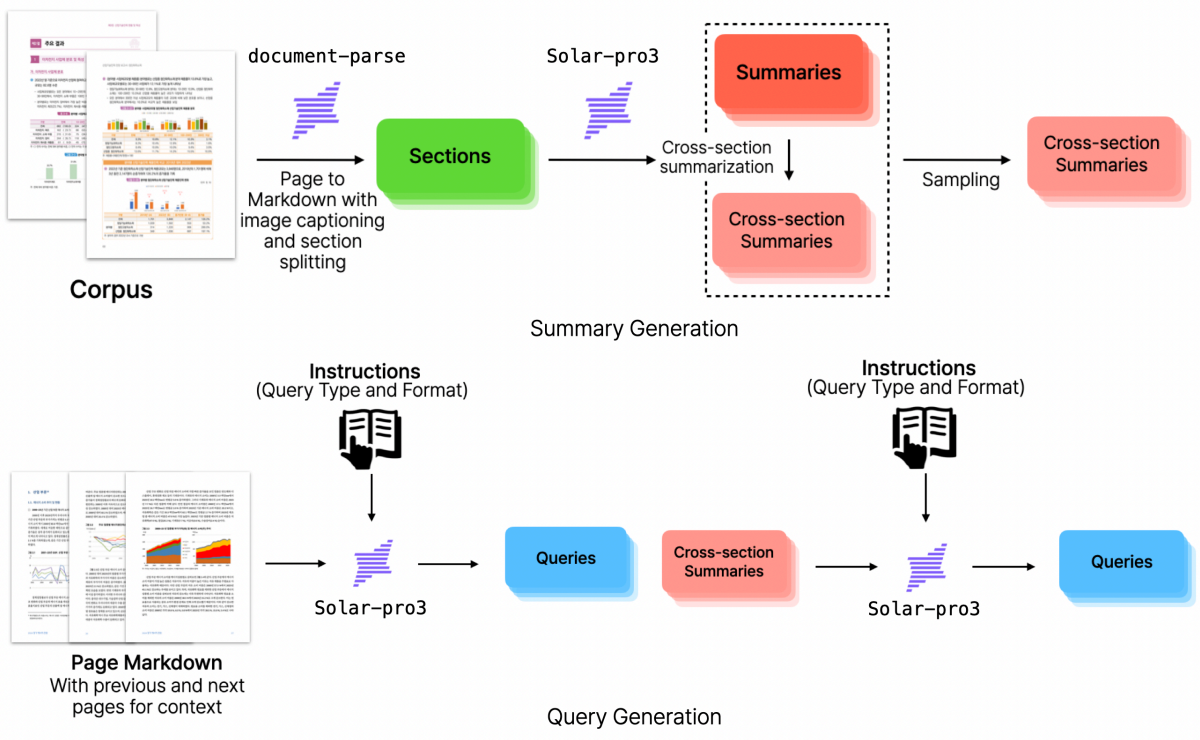


Figure 2: Overview of the KoViDoRe data generation pipeline. Queries are generated through both summary-based and context-based strategies. Summary representations capture global document relationships, while context-based generation focuses on local content. The pipeline further includes filtering and manual verification to ensure query quality and reliable relevance mapping.

also leverages publicly accessible Korean documents, we focus on authentic document sources rather than synthetic or simplified formats. This allows the benchmark to reflect real-world document characteristics, including complex layouts, tables, figures, and multi-column structures.

3.2 Document Processing and Parsing

We first split each PDF document into individual pages and perform processing at the page level. This allows us to treat each page as a fundamental unit while preserving the document structure. To extract structured representations from each page, we employ the Upstage Document Parse[2], which is effective for parsing structurally complex Korean documents while preserving layout-aware information. For every page, textual content and visual elements are identified and organized into semantically meaningful sections. Specifically, the parser provides both page-level markdown and element-level markdown for each page. The page-level markdown offers a unified view of the entire page content, while the element-level markdown decomposes the page into fine-grained components such as text blocks, tables, figures, charts, and diagrams. In addition to structural extraction, the document parser provides captions for visual elements such as figures, charts, and diagrams. These captions offer semantic descriptions of visual content, enabling better understanding of non-textual information during downstream processing. By combining page-level and element-level markdown with captioned visual elements, our preprocessing pipeline preserves document-level structure while enabling fine-grained and semantically enriched access to page content.

3.3 Query Generation

To construct a diverse and scalable set of queries, we adopt a synthetic query generation pipeline based on reasoning-oriented language model, Solar-Pro3[3], which supports strong Korean language understanding and is well-suited for generating complex Korean queries that reflect multiple information needs. The model generates queries by conditioning on document content, including both textual and visual information extracted during preprocessing.

[2]document-parse-251217

[3]solar-pro3-260126

Summary-based Query Generation Before query generation, we first construct intermediate summaries to better expose document structure and cross-page relationships. Specifically, we generate two types of summaries. The first type consists of single-section summaries that describe individual sections. The second type consists of cross-section summaries, which are constructed by randomly sampling multiple single-section summaries (e.g., 3, 5, or 7 sections) and synthesizing them to capture cross-sectional relationships. These summaries provide a higher-level abstraction of document content, allowing the generation process to capture relationships that are not easily observable from isolated pages. Queries generated from summaries therefore tend to reflect more global information needs and often require reasoning across multiple sections or pages.

Context-based Query Generation In addition to summary-based generation, we also generate queries directly from local document context. In this setting, the model is prompted using page-level or local multi-page context windows, enabling it to produce queries grounded in nearby content. This complementary route helps capture more localized information needs and preserves natural query patterns tied to specific document regions. Combining both summary-based and context-based queries allows the dataset to cover a broader spectrum of retrieval scenarios.

Diversity Control To promote diversity, we adopt the query formulation scheme introduced in ViDoRe V3 (Loison et al., 2026). Specifically, we control the query format and type during generation, enabling the construction of a diverse set of queries with different structural patterns and information needs. As a result, the dataset includes various query types such as multi-hop reasoning, numerical comparison, and aggregation-based queries, where a single query may exhibit multiple types simultaneously. The full set of query types and formats, along with their definitions, is summarized in Table 5 and Table 6.

3.4 Relevance Mapping and Filtering

To construct reliable query-document relevance annotations, we incorporate relevance mapping into multiple stages of the pipeline. During query generation, the model is provided with document content in markdown format, including both page-level and element-level representations, and queries are generated based on specific sections or combinations of sections, implicitly capturing initial relevance signals. After query generation, we perform an additional relevance mapping step at the page level by evaluating each query against candidate document pages to identify supporting evidence.

To improve annotation quality while reducing manual effort, we apply both consistency-based and rule-based filtering. Relevance signals from the generation stage are compared with those from the additional mapping stage, and only consistent pairs are retained. While this process may discard some challenging cases where relevant pages are difficult to identify during mapping, we prioritize annotation reliability over coverage, as relevance judgments can be inherently ambiguous in mul-

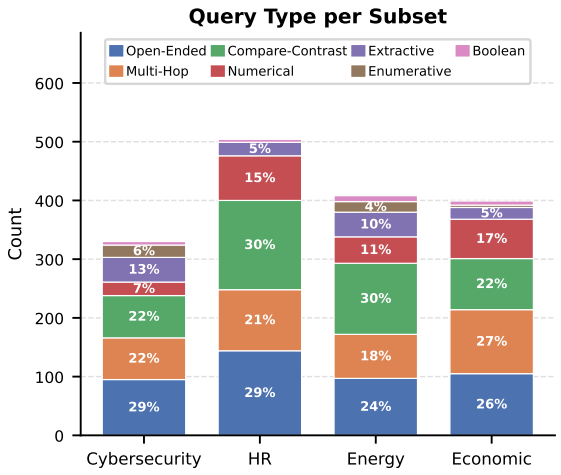


Figure 3: Distribution of query types across subsets. Note that query types are not mutually exclusive, accounting for the multi-faceted nature of complex information needs.

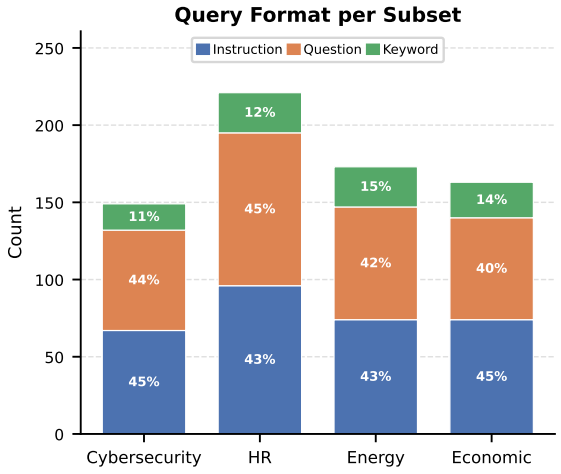


Figure 4: Distribution of query formats across subsets.

Subset	#Docs	#Pages	#Queries	#Qrels	Avg. Pages / Query
Cybersecurity	17	1,150	149	409	2.74
HR	9	2,109	221	726	3.30
Energy	11	1,993	173	525	3.03
Economic	20	1,477	163	413	2.55
Total	57	6,729	706	2,073	2.94

Table 1: Statistics of the KoViDoRe benchmark across domain-specific subsets. Avg. Pages / Query denotes the average number of relevant pages per query.

timodal and multi-page settings. In addition, we remove low-quality queries such as those with excessive keyword enumeration or those that directly reveal answer content from the document, resulting in a reduced but more reliable set of keyword-based queries. For the benchmark, we further incorporate human verification, where annotators perform a final review of query-page relevance annotations and refine queries through rephrasing when they are unnatural or ambiguous. This process balances automatic filtering and human verification to produce high-quality annotations.

3.5 Dataset Statistics

Table 1 summarizes the overall statistics of the benchmark across its four domain-specific subsets. Figure 3 and Figure 4 further illustrate the distributions of query types and query formats. The benchmark consists of 57 documents and 6,729 pages, with a total of 706 queries and 2,073 relevance annotations. The four domain-specific subsets exhibit notable differences in scale and structure. For example, the HR and Energy subsets contain a larger number of pages per query, suggesting that queries in these domains often require aggregating information from multiple pages. As shown in Figure 3 and Figure 4, the dataset contains a diverse set of query types and formats across all subsets. Multi-hop, open-ended, and comparison-based queries appear frequently, while question- and instruction-style queries are more common than keyword queries.

4 Experiments

4.1 Experimental Setup

We evaluate Korean visual document retrieval as a ranking task over document pages. Given a query q and a collection of document pages $\mathcal{D}$, the goal is to retrieve and rank pages that are relevant to the query. Each document is represented as a set of pages containing both textual and visual content. Queries are written in Korean and reflect diverse information needs grounded in real-world documents. Relevance is defined at the page level with graded labels: a page is labeled as fully relevant (2) if it contains sufficient information to answer the query, and partially relevant (1) if it provides supporting evidence. The dataset follows the BEIR-style evaluation framework with a corpus, queries, and relevance judgments (qrels) (Thakur et al., 2021). We report performance using nDCG@10. We evaluate a range of multimodal retrieval models on the KoViDoRe benchmark, covering small ($<$1B), medium (1B–4B), and large ($\geq$4B) models. The evaluated models include CLIP-based encoders (Radford et al., 2021; Zhai et al., 2023; Koukounas et al., 2024), late-interaction retrieval models such as ColPali and ColQwen (Faysse et al., 2024; Nomic Team, 2025; Huang and Tan, 2025), and recent multimodal embedding models (Jiang et al., 2024; Günther et al., 2025; Li et al., 2026). Evaluation is conducted across four domains: Cybersecurity, Energy, Economic, and Human Resources.

We conduct evaluation using the MTEB (Massive Text Embedding Benchmark) (Muennighoff et al., 2022), adapted to support multimodal retrieval on the KoViDoRe dataset. This provides a standardized and reproducible evaluation pipeline across all models. We build separate retrieval indices for each domain subset and each query is evaluated only against the pages within its corresponding domain.

4.2 Main Results

Table 2 presents the performance of all evaluated models on the benchmark. Overall, we observe a clear performance gap between model scales. Small models perform poorly across all domains, often failing to retrieve relevant pages. Medium-scale models show improved performance, particularly those based on late-interaction architectures. Larger models generally achieve higher scores, although performance gains are not uniform across domains. Among all models, `jina-embeddings-v4`

Model	Params	Cyber	Energy	Economic	HR	Avg.
Small Models (<1B parameters)						
openai/clip-vit-base-patch16♠	151M	4.1	0.8	0.0	0.6	1.4
vidore/colSmol-256M♦	256M	19.7	9.5	1.0	1.1	7.8
vidore/colSmol-500M♦	500M	26.2	9.9	0.6	0.9	9.4
jinaai/jina-clip-v2♠	865M	20.4	11.3	0.2	3.1	8.8
google/siglip-so400m-patch14-384♠	878M	15.3	5.3	1.3	1.1	5.8
Medium Models (1B–4B parameters)						
Qwen/Qwen3-VL-Embedding-2B♣	2.0B	61.3	40.6	15.3	18.7	34.0
vidore/colqwen2-v1.0♦	2.2B	53.3	42.0	8.0	14.7	29.5
vidore/colpali-v1.1♦	2.9B	31.9	18.2	3.0	6.0	14.8
vidore/colpali-v1.2♦	2.9B	33.2	16.4	2.1	4.5	14.1
vidore/colpali-v1.3♦	2.9B	34.7	20.6	1.6	6.2	15.8
ApsaraStackMaaS/EvoQwen2.5-VL-Retriever-3B-v1♦	3.0B	41.4	31.5	6.3	11.3	22.6
nomic-ai/colnomic-embed-multimodal-3b♦	3.0B	47.4	44.2	10.5	32.9	33.7
vidore/colqwen2.5-v0.2♦	3.0B	43.9	44.3	3.9	13.5	26.4
jinaai/jina-embeddings-v4♣	3.8B	<u>77.6</u>	**67.7**	**24.5**	**50.1**	**55.0**
Large Models (≥4B parameters)						
TomoroAI/tomoro-colqwen3-embed-4b♦	4.0B	55.3	31.0	9.1	10.1	26.4
eagerworks/eager-embed-v1♣	4.0B	51.5	32.7	5.4	7.0	24.2
TIGER-Lab/VLM2Vec-Full♣	4.2B	9.8	3.2	1.3	1.3	3.9
ApsaraStackMaaS/EvoQwen2.5-VL-Retriever-7B-v1♦	7.0B	66.0	55.4	12.1	26.4	40.0
nomic-ai/colnomic-embed-multimodal-7b♦	7.0B	69.6	59.5	12.4	33.3	43.7
Qwen/Qwen3-VL-Embedding-8B♣	8.0B	**77.8**	<u>63.2</u>	<u>23.4</u>	<u>37.4</u>	<u>50.4</u>
TomoroAI/tomoro-colqwen3-embed-8b♦	8.0B	73.7	58.5	16.3	26.5	43.8

Table 2: Performance comparison on KoViDoRe benchmark (nDCG@10, %). **Bold** indicates the best score; <u>underline</u> indicates the second-best score. ♠: CLIP-based, ♦: late-interaction, ♣: single-vector models. Cyber: Cybersecurity, HR: Human Resources.

achieves the best overall performance, significantly outperforming other models across all domains. This suggests that retrieval effectiveness is influenced not only by model scale, but also by factors such as training objective and data composition.

4.3 Analysis

Despite improvements from larger models, performance remains limited across all domains. In particular, domains such as Economic and Human Resources consistently show lower scores, indicating that retrieving relevant information in these settings is especially challenging. This is likely due to more complex document structures and the presence of information distributed across multiple pages.

We also observe that even strong retrieval models achieve relatively limited performance on several subsets of KoViDoRe. This suggests that the benchmark introduces additional challenges beyond conventional document retrieval settings, including complex layouts, structured visual content, and information distributed across multiple pages. In particular, queries associated with multiple relevant pages remain difficult for existing models, indicating that effectively retrieving and aggregating distributed document evidence is still a challenging problem in Korean visual document retrieval.

Motivated by this limitation, we further investigate whether training on Korean-specific data can improve retrieval performance, which we explore in the following subsection.

4.4 Ko-VDR Train Public

Dataset Construction To address the limitations identified in the previous section, we curate a large-scale training dataset, **Ko-VDR Train Public**, using the same data generation pipeline. The dataset consists of query-page pairs derived from Korean visual documents and includes a total of **310,226** query-page pairs. To ensure data quality, we apply both consistency-based and rule-based filtering. The consistency-based filtering retains only query-page pairs where relevance signals from the query generation stage and the additional relevance mapping stage agree. In addition, we apply rule-based filtering to remove low-quality queries, including those with excessive keyword enumeration and those that directly reveal answer content from

Model	Params	Cyber	Energy	Economic	HR	Avg.
vidore/colSmol-500M	500M	26.2	9.9	0.6	0.9	9.4
+ Ko-VDR Train Public		39.4	35.0	14.4	18.7	26.9
vidore/colqwen2-v1.0	2.2B	53.3	42.0	8.0	14.7	29.5
+ Ko-VDR Train Public		75.6	67.6	18.3	49.6	52.8
jinaai/jina-embeddings-v4	3.8B	77.6	**67.7**	**24.5**	**50.1**	**55.0**
TomoroAI/tomoro-colqwen3-embed-4b	4.0B	55.3	31.0	9.1	10.1	26.4
Qwen/Qwen3-VL-Embedding-8B	8.0B	**77.8**	63.2	23.4	37.4	50.4

Table 3: Comparison with representative retrieval baselines on KoViDoRe (nDCG@10, %). **Bold** indicates the best score; underline indicates the second-best score. Cyber: Cybersecurity, HR: Human Resources.

the document. This helps eliminate trivial or overly extractive cases and improves the robustness of the data. Unlike the benchmark construction process, we do not perform additional human verification for the training dataset to maintain scalability.

Training Setup We fine-tune two late-interaction retrieval models, `colSmol-500M` and `colqwen2-v1.0`, using the `colpali_engine` framework. Training is conducted on 2× NVIDIA B200 GPUs using BF16. We use a batch size of 128 per device and train for 3 epochs. In addition to Ko-VDR Train Public, we mix in a private Korean visual QA dataset containing TableVQA- and FigureVQA-style supervision (Kim et al., 2024; Kahou et al., 2017). This additional dataset complements the retrieval objective by providing stronger supervision for structured visual understanding, particularly for tables and figures.

Results Table 3 shows that fine-tuning on our dataset consistently improves performance across both `colSmol-500M` and `colqwen2-v1.0`. The smaller `colSmol-500M` model achieves substantial gains across all domains, demonstrating the effectiveness of the proposed training data even for lightweight models. For `colqwen2-v1.0`, fine-tuning leads to significant performance improvements, achieving competitive results with strong multimodal embedding models and surpassing `Qwen3-VL-Embedding-8B` despite being smaller in scale. These results highlight that training on Korean-specific data can substantially improve retrieval performance and enable smaller models to compete with larger counterparts.

4.5 Interpretability

To better understand model behavior, we visualize query-to-document similarity using heatmaps for the fine-tuned `colqwen2-v1.0` model. In Figure 5,

Query: 독일 현물시장 참가자 수 감소가 **선물시장** 비중 확대와 시장 구조 분화와 관련이 있나요?

Figure 5: Similarity map on a document example.

the model assigns high similarity to the term "선물시장," indicating that it correctly focuses on the key textual evidence relevant to the query. This suggests that the model is able to identify and attend to query-relevant terms in Korean visual documents. Additional examples are provided in Figure 15.

5 Ablation Study

5.1 Effect of the Number of Relevant Pages

To analyze how retrieval performance varies with query complexity, we group queries by the number of relevant pages and report nDCG@10 for each group. Figure 6 shows the results for `Qwen3-VL-Embedding-2B` and `Qwen3-VL-Embedding-8B`. We observe a general trend where performance tends to decrease as the number of relevant pages increases. For both models, performance is highest when a query is

Model	Cyber	Energy	Economic	HR	Avg.
vidore/colqwen2-v1.0	53.3	42.0	8.0	14.7	29.5
+ Private Only	70.3	58.8	**18.7**	37.7	46.4
+ Public Only	<u>75.4</u>	<u>66.9</u>	16.5	<u>49.3</u>	<u>52.0</u>
+ **Private + Public**	**75.6**	**67.6**	<u>18.3</u>	**49.6**	**52.8**

Table 4: Effect of training data composition on `vidore/colqwen2-v1.0` (nDCG@10, %). **Bold** indicates the best score; <u>underline</u> indicates the second-best score. Cyber: Cybersecurity, HR: Human Resources.

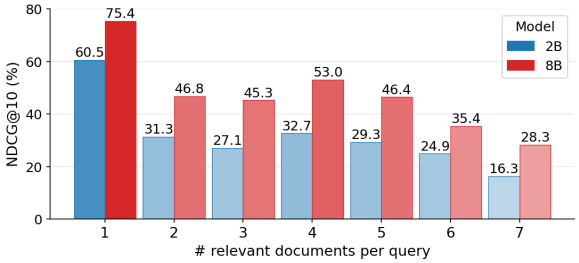


Figure 6: nDCG@10 by the number of relevant pages per query. Performance tends to decrease as the number of relevant pages increases.

associated with a single relevant page, and generally declines as more relevant pages are required, although the decrease is not strictly monotonic. This suggests that queries requiring evidence from multiple pages are more challenging, as models must retrieve and integrate information distributed across different parts of a document. While `Qwen3-VL-Embedding-8B` consistently outperforms `Qwen3-VL-Embedding-2B` across all groups, both exhibit similar trends, indicating that increasing model capacity alone does not fully mitigate the challenges of multi-page retrieval. These results highlight that KoViDoRe captures the increased difficulty of queries with broader information needs, providing a more realistic evaluation setting in which retrieval systems must aggregate evidence across multiple pages.

5.2 Effect of Training Data Composition

We analyze the effect of training data composition using `vidore/colqwen2-v1.0`. Following the training setup described in Section 4.4, we keep all training configurations fixed and vary only the composition of the training data. As shown in Table 4, training with private data alone improves performance over the base model, indicating the benefit of Korean-specific supervision. Training with public data yields larger improvements across most subsets, suggesting that broader data coverage and diversity contribute significantly to retrieval performance on Korean visual documents. When both private and public data are combined, the model achieves the best overall performance across most subsets, demonstrating that Korean-specific supervision and large-scale public training data are complementary. In particular, combining both datasets consistently improves performance in Cybersecurity, Energy, and HR, leading to the highest average performance overall. Interestingly, the Economic subset exhibits a different trend, where training with private data alone achieves the highest performance. We hypothesize that this is because many queries in the Economic subset require identifying and interpreting complex multi-column tables distributed across document pages. Since the private dataset includes TableVQA-style supervision, it likely provides stronger training signals for structured table understanding, resulting in larger gains on table-heavy economic documents.

6 Conclusion

We introduced KoViDoRe, a benchmark for Korean visual document retrieval. Unlike conventional benchmarks that primarily focus on queries answerable from a single page, KoViDoRe emphasizes queries that require retrieving and integrating information distributed across multiple pages. We constructed the dataset from publicly available Korean documents with diverse layouts and developed a LLM-based multi-stage pipeline with human verification. Through extensive evaluation, we showed that current multimodal retrieval models struggle to effectively handle Korean visual document retrieval, particularly in scenarios involving structured content and diverse query types. To address this limitation, we further curated Ko-VDR Train Public, a large-scale training dataset designed for Korean visual document retrieval. Our experiments demonstrate that training on Korean-specific data improves retrieval performance, highlighting the importance of language-specific training resources. We hope that KoViDoRe and Ko-VDR Train Public will facilitate future research on Korean visual documents retrieval.

Limitations

Despite the contributions of this work, several limitations remain. First, our query generation relies on parsed markdown representations and image captions rather than raw visual inputs. While this design enables scalable and reproducible data generation, it may not fully preserve fine-grained visual information such as layout, color, or chart-specific patterns. As a result, information loss may occur in visually intensive documents, potentially affecting query quality. As future work, we plan to incorporate vision-language models (VLMs) into the query generation process to better capture visual information and reduce such information loss. Second, while our relevance mapping process reduces manual annotation effort through consistency-based filtering, it may still introduce noise due to imperfect alignment between generation and mapping stages. In addition, although the private Korean VQA dataset used in our training experiments contributes to performance improvements, it cannot be publicly released due to licensing restrictions. As a result, the fully reproducible training setup is limited to the publicly available components. Finally, our experiments focus on evaluating existing retrieval models, and we do not propose new model architectures specifically designed for Korean visual document retrieval. Future work may explore model designs and training strategies better suited for handling structured and visually rich documents, including approaches that directly incorporate visual inputs during query generation.

Acknowledgments

This work was supported by the IITP(Institute of Information & Communications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT) (IITP-2026-RS-2024-00438239).

References

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. 2025. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16776–16809, Vienna, Austria. Association for Computational Linguistics.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. 2024. M3docrag: Multimodal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*.

Kuicai Dong, Yujing Chang, Derrick Goh Xin Deik, Dexun Li, Ruiming Tang, and Yong Liu. 2025. MMDocIR: Benchmarking multimodal retrieval for long documents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30971–31005, Suzhou, China. Association for Computational Linguistics.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, and 1 others. 2025. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 531–550.

Xin Huang and Kye Min Tan. 2025. Beyond text: Unlocking true multimodal, end-to-end rag with tomoro colqwen3.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2024. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. 2024. Tablevqa-bench: A visual question answering benchmark on multiple table domains. *arXiv preprint arXiv:2404.19205*.

Andreas Koukounas, Georgios Mastrapas, Sedigheh Eslami, Bo Wang, Mohammad Kalim Akram, Michael Günther, Isabelle Mohr, Saba Sturua, Nan Wang, and Han Xiao. 2024. jina-clip-v2: Multilingual multimodal embeddings for text and images. *arXiv preprint arXiv:2412.08802*.

Jaehoon Lee, Sohyun Kim, Wanggeun Park, Geon Lee, Seungkyung Kim, and Minyoung Lee. 2025. Sds kopub vdr: A benchmark dataset for visual document retrieval in korean public documents. *arXiv preprint arXiv:2511.04910*.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Chen Keqin, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun

Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*.

António Loison, Quentin Macé, Antoine Edy, Victor Xing, Tom Balough, Gabriel Moreira, Bo Liu, Manuel Faysse, Céline Hudelot, and Gautier Viaud. 2026. Vidore v3: A comprehensive evaluation of retrieval augmented generation in complex real-world scenarios. *arXiv preprint arXiv:2601.08620*.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6492–6505, Miami, Florida, USA. Association for Computational Linguistics.

Quentin Macé, António Loison, and Manuel Faysse. 2025. Vidore benchmark v2: Raising the bar for visual retrieval. *arXiv preprint arXiv:2505.17166*.

Gabriel de Souza P Moreira, Ronay Ak, Mengyao Xu, Oliver Holworthy, Benedikt Schifferer, Zhiding Yu, Yauhen Babakhin, Radek Osmulski, Jiarui Cai, Ryan Chesler, and 1 others. 2026. Nemotron colembed v2: Top-performing late interaction embedding models for visual document retrieval. *arXiv preprint arXiv:2602.03992*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Nomic Team. 2025. Nomic embed multimodal: Interleaved text, image, and screenshots for visual document retrieval.

Radek Osmulski, Gabriel de Souza P Moreira, Ronay Ak, Mengyao Xu, Benedikt Schifferer, and Even Oldridge. 2025. Miracl-vision: A large, multilingual, visual document retrieval benchmark. *arXiv preprint arXiv:2505.11651*.

Xiangyu Peng, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, and Chien-Sheng Wu. 2025. Unidocbench: A unified benchmark for document-centric multimodal rag. *arXiv preprint arXiv:2510.03663*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Connor Shorten, Augustas Skaburskas, Daniel M Jones, Charles Pierse, Roberto Esposito, John Trengrove, Etienne Dilocker, and Bob van Luijt. 2026. Irpapers: A visual document benchmark for scientific retrieval and question answering. *arXiv preprint arXiv:2602.17687*.

Shezheng Song, Xiaopeng Li, Shasha Li, Shan Zhao, Jie Yu, Jun Ma, Xiaoguang Mao, Weimin Zhang, and Meng Wang. 2025. How to bridge the gap between modalities: Survey on multimodal large language model. *IEEE Transactions on Knowledge and Data Engineering*, 37(9):5311–5329.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. 2025. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9124–9145.

Navve Wasserman, Roi Pony, Oshri Naparstek, Adi Raz Goldfarb, Eli Schwartz, Udi Barzelay, and Leonid Karlinsky. 2025. REAL-MM-RAG: A real-world multi-modal retrieval benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31660–31683, Vienna, Austria. Association for Computational Linguistics.

Zilin Xiao, Qi Ma, Mengting Gu, Chun-cheng Jason Chen, Xintao Chen, Vicente Ordonez, and Vijai Mohan. 2025. Metaembed: Scaling multimodal retrieval at test-time with flexible late interaction. *arXiv preprint arXiv:2509.18095*.

Yibo Yan, Jiahao Huo, Guanbo Feng, Mingdong Ou, Yi Cao, Xin Zou, Shuliang Liu, Yuanhuiyi Lyu, Yu Huang, Jungang Li, and 1 others. 2026. Unlocking multimodal document intelligence: From current triumphs to future frontiers of visual document retrieval. *arXiv preprint arXiv:2602.19961*.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and 1 others. 2024. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

A Appendix

This appendix provides supplementary materials for transparency and reproducibility. We include a detailed comparison with the SDS KoPub VDR benchmark, the technical rationale for our model selections, and formal category definitions used in query generation. Additionally, we provide the complete set of prompt templates for key pipeline stages, representative query-page examples, and comprehensive document metadata for all source collections included in KoViDoRe.

A.1 Comparison to SDS KoPub VDR

While SDS KoPub VDR (Lee et al., 2025) represents an important benchmark for evaluating Korean visual document retrieval, KoViDoRe introduces a different task formulation that emphasizes more complex retrieval scenarios. The primary distinction lies in the query-to-document mapping: SDS KoPub VDR is largely designed for single-page retrieval, where each query is mapped to a single relevant page. In contrast, KoViDoRe explicitly focuses on multi-page evidence aggregation, with each query associated with an average of 2.94 relevant pages. This shift from single-page matching to multi-page evidence gathering aligns with realistic enterprise search scenarios, where information is often distributed across multiple pages and no single page alone is sufficient to satisfy the query. As shown in Table 1, KoViDoRe's queries frequently require synthesizing information from diverse document regions, such as comparing financial trends or summarizing policies spread throughout a report. By providing a higher ratio of relevant pages per query—reaching up to 3.30 in the HR subset—KoViDoRe complements existing resources by evaluating a model's ability to retrieve information from multiple pages to satisfy the diverse informational needs embedded in a single query.

A.2 Document Parsing and Query Generation Models

Upstage Document Parse We use Upstage Document Parse as the document parsing backend in our pipeline. The model provides layout-aware parsing of visually rich documents, extracting both page-level and element-level markdown representations, and decomposing each page into structured components such as text blocks, tables, figures, charts, and diagrams. It also generates captions for visual elements, enabling semantic interpretation of non-textual content. According to publicly reported results on DP-Bench, the model achieves strong performance in preserving document structure, and is designed to handle complex document layouts. We choose this model as it is effective for parsing structurally complex Korean documents while preserving layout-aware information, which is critical for downstream query generation and relevance mapping.

Solar-Pro3 For query generation, we use Solar-Pro3 as the underlying language model. Solar-Pro3 is a reasoning-oriented language model designed to support structured and context-aware generation. According to publicly available reports, it demonstrates strong performance in Korean language understanding and instruction-following tasks. Given document content in markdown format, the model generates queries conditioned on both textual and visual information extracted during preprocessing. We choose Solar-Pro3 as it is well-suited for generating complex Korean queries that require multi-step reasoning, which aligns with the objective of constructing realistic and challenging retrieval scenarios in KoViDoRe.

A.3 Query Category Definitions

Table 5 and Table 6 define the query type and query format categories used in our dataset construction pipeline.

A.4 Prompt Templates

Figures 9, 10, 11, and 12 present the prompt templates used for summary generation, relevance mapping, and query generation.

A.5 Example Query-Page Pairs

Figure 13 and Figure 14 show representative examples of query-page pairs from different subsets of KoViDoRe.

A.6 Document Metadata

Table 7 lists the metadata of the document collections used in KoViDoRe, while Table 8 presents the metadata of the documents used in Ko-VDR Train Public. Both tables include document titles, providers, page counts, and license information.

A.7 Distribution of Relevant Pages per Query

Figure 7 shows the distribution of queries with respect to the number of annotated relevant pages.

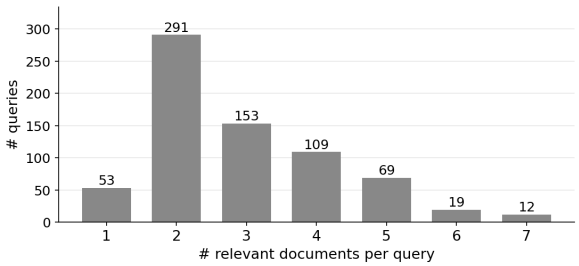


Figure 7: Distribution of queries by the number of relevant pages.

Most queries are associated with two or three relevant pages, while queries requiring a larger number of relevant pages are less frequent. This distribution reflects a realistic setting where information needs vary in complexity, with a substantial portion of queries requiring aggregation across multiple pages. At the same time, the presence of queries with a higher number of relevant pages supports the analysis in Section 5.1, demonstrating that KoViDoRe includes challenging cases that require broader evidence aggregation.

Category	Definition
open-ended	A query requiring synthesis and explanation of information. The answer must integrate multiple concepts into a coherent narrative rather than citing a single fact.
compare-contrast	A query requiring identification and articulation of similarities and/or differences between two or more entities, concepts, or topics.
enumerative	A query requesting a complete list of items that meet specific criteria.
numerical	A query expecting a numerical value, obtained either by direct extraction or calculation.
boolean	A query expecting a yes/no answer, potentially requiring reasoning over extracted information.
extractive	A query answerable by directly citing a specific fact or piece of information from the documents.
multi-hop	A query requiring information retrieval from multiple distinct sources or sections, which must then be combined to produce a complete answer.

Table 5: Query type categories and their definitions.

Category	Definition
question	A query presented in the form of a direct question, seeking specific information or clarification.
instruction	A query framed as a directive or command, requesting the model to perform a specific task or provide information in a particular manner.
keyword	A query consisting of noun phrases and keywords only, WITHOUT forming a complete sentence. No verbs, no question words, no sentence endings. Mimics how users type into search engines: fragmented, concise, noun-centric terms separated by spaces.

Table 6: Query format categories and their definitions.

```
You are a document analysis expert. Your role is to analyze a specific section of a document
within the context of the entire page and summarize its core content in Korean.

## Context and Task:
- You are provided with the full page content in markdown format.
- You are also provided with a specific section's data, which may include text segments,
coordinates, and metadata.
- Your task is to summarize the core information of the **Target Section ** while using the
**Full Page** as contextual background to ensure accuracy and completeness.

## Summary Principles:
- Capture all key information, concepts, and data specifically related to the Target Section.
- If the Target Section refers to tables, charts, or graphs present in the Full Page,
describe their content and significance.
- Ensure that numerical data, statistics, and important figures from the section are
strictly included.
- The summary should be 5-7 sentences long—concise, yet minimizing any loss of information.
- A reader should be able to understand the core message of the specific section solely by
reading the summary.

## Output Format:
- Output only the summary written in Korean.
- Do not provide any introductory remarks, preambles, or additional explanations.

## Instructions:
Please summarize the Target Section based on the Full Page context.

### Full Page Markdown:
{{ markdown }}

### Target Section:
{{ elements }}
```

Figure 8: Prompt for generating single-section summaries based on full-page context

```
You are a document synthesis and integration expert. Your role is to analyze a set of
fragmented summaries extracted from different pages of a single document and synthesize them
into a coherent, comprehensive cross-section summary in Korean.

## Context and Task:
- You are provided with a **Combined Context**, which consists of multiple summaries derived
from randomly sampled sections of a document.
- The context is listed with page numbers to indicate where each information originates.
- Your task is to generate a **Cross-Section Summary** that integrates these dispersed
pieces of information into a unified narrative.
- You must identify logical connections, thematic consistency, or causal relationships
between the sections, even if the page numbers are not consecutive.

## Summary Principles:
- **Integration over Listing:** Do not simply list the summaries one by one. Instead, weave
them together to explain "what this document is discussing" based on the available evidence.
- **Contextual Flow:** Use the page numbers to infer the structure (e.g., "The document
introduces [Topic] on Page 2 and later elaborates on [Specific Detail] on Page 15").
- **Handling Gaps:** Acknowledge that the information is sampled. If sections seem unrelated,
describe them as distinct aspects covered within the document.
- **Accuracy:** Strictly adhere to the provided content. Do not hallucinate information not
present in the input snippets.
- **Length & Tone:** Write a professional, dense paragraph (7-10 sentences). Use a formal
and objective tone.

## Output Format:
- Output only the summary written in Korean.
- Do not provide any introductory remarks, preambles, or additional explanations.

## Instructions:
Please summarize the core message based on the provided Combined Context below.

### Combined Context:
<sections>
{% for summary in single_section_summary %}
<section index="{{ loop.index0 }}">
{{ summary }}
</section>
{% endfor %}
</sections>
```

Figure 9: Prompt for cross-section summarization using aggregated section summaries

```
You are a strategic Document Relevance Auditor. Your goal is to identify pages that provide
either a "Complete Answer" or "Essential Building Blocks" for a query.

## Task
Evaluate the relevance of each document page. You must distinguish between "Noisy/Empty
pages" and "Partial but Crucial data pages."

## Query
{{ query }}

## Documents
<documents>
{% for doc in markdown %}
<document index="{{ loop.index0 }}">
{{ doc }}
</document>
{% endfor %}
</documents>

## Scoring Criteria (Balanced Evidence-Based)

- **2 (FULLY_RELEVANT)**: The page contains an explicit, direct, and complete answer to all
parts of the query.
- **1 (CRITICALLY_RELEVANT)**: The page contains specific, substantive facts or data
required to answer *at least one part* of a multi-part query.
    - (e.g., If the query asks for "A and B comparison" and the page has detailed data on
    "A", it is a CRITICAL building block, even if "B" is missing.)
    - (e.g., Detailed statistics, specific policy names, or factual descriptions that would
    form part of the final answer.)
- **0 (IRRELEVANT)**: The page provides no substantive value. This includes:
    - **Pure Navigation**: Tables of Contents or cover pages with only titles/page numbers.
    - **Off-Topic**: Content that doesn't address any specific component of the query.
    - **Vague Mentions**: Just mentioning a keyword without any descriptive facts or data.

## Critical Instructions
1. **The Building Block Rule**: Do not reject a page just because it is incomplete. If it
provides a "Hard Fact" (e.g., China's specific Metaverse policy) that is part of the query's
scope, assign 1.
2. **Substance Over Format**: A table or a list of policies is highly relevant if it contains
the "What/How/When" of the subject, even if it doesn't "compare" it for you.
3. **Anti-Hallucination**: While being more inclusive of partial data, still score 0 if the
page requires you to "guess" the information. The data must be explicitly written.

## Reasoning Requirements (Thinking Process)
For each page, explain your judgment in **KOREAN**:
- **Partial match check**: Does this page cover at least one specific component of the query?
- **Fact density**: Does it provide concrete data/facts, or just general mentions?
- **Role in Answer**: How does this information help in constructing the final response?

## Output Format
Return your assessment with:
1. `reasoning`: A detailed **KOREAN** explanation for each page.
2. `relevance_scores`: A list of integer scores (0, 1, or 2).
```

Figure 10: Prompt for relevance mapping by identifying fully relevant and critically relevant pages

```
You are an expert in creating challenging datasets for Vision Document Retrieval (VDR).
Your goal is to generate a **highly specific Korean search query** that acts as a realistic
user prompt for retrieving information from a large corpus.

### 1. Document Context
You are provided with two types of summaries:

#### 1.1 Single-Section Summaries
Each <single_section_summary> tag contains a summary of a **specific section/page**,
identified by its **actual page number** (page).

<single_section_summaries>
{% for summary in single_section_summary %}
<single_section_summary index="{{ loop.index0 }}">
{{ summary }}
</single_section_summary>
{% endfor %}
</single_section_summaries>

#### 1.2 Cross-Section Summary
The following is a **synthesized summary** that integrates information across all the
sections above. Use this to understand the overall narrative and relationships between
different sections.

<cross_section_summary>
{{ cross_section_summary }}
</cross_section_summary>

**How to use these summaries**:
* Use **single-section summaries** to identify specific facts, entities, and details located
on each page.
* Use **cross-section summary** to understand how information connects across pages and to
identify synthesis opportunities.
* Your query should require combining specific details from multiple sections (identified via
single-section summaries) in a way that reflects the cross-section relationships (identified
via cross-section summary).

### 2. Task Requirements
You must generate a structured output containing the rationale and the query itself based on
the following specifications:

* **Query Type**: {{ query_type }} ({{ query_type_definition }})
* **Query Format**: {{ query_format }} ({{ query_format_definition }})

### 3. Critical Constraints for Realistic Retrieval

#### Rule 1: NO Artificial Location References
* **Strictly Forbidden**: "2페이지에서...", "다음 장에 있는...", "첫 번째 문서의...", "위에서
언급된..."
* **Reason**: The user queries the entire database and does not know the document order or
page numbers.
* **Alternative**: Use **Section Headers, Table Captions, or Unique Keywords** found in the
text.
  * Bad: "2페이지에 있는 표를 요약해."
  * Good: "'2024년 재무 하이라이트' 표를 요약해."

#### Rule 2: Implicit Multi-Page Synthesis
* The query must require information scattered across multiple pages, but **without
explicitly stating so**.
* **Strategy**: Identify **Entity A** on one page and **Entity B** on another, then ask about
their relationship.

#### Rule 3: Entity-Grounded Specificity
* Avoid generic queries like "에너지 정책을 분석해줘."
* Include specific entities found in the text: **Dates, Company Names, Regulations (e.g.,
ISO-27001), Project Codes, Policy Names, or Program Names**.
* However, do NOT include exact numerical values (see Rule 5).

#### Rule 4: Single Natural Query
* The query **MUST be a single unit** appropriate to its format (one question, one
instruction, or one keyword cluster).
* **Strictly Forbidden Patterns**:
  * Multiple sentences: "~입니다. ~해주세요."
```

  * Instruction suffixes: "단, ~를 기준으로 답변하시오."
  * Explicit output format requests: "~를 근거로 제시하시오.", "~를 나열하시오."
  * Conditional clauses at the end: "단, ~를 구분하여 제공해야 합니다."
* **Examples**:
  * Bad: "2021년과 2022년 상승률을 비교하시오. 단, 수도권과 지방을 구분하여 제시하시오."
  * Good: "2021년과 2022년 수도권 및 지방의 주택 매매가격 상승률은 어떻게 달랐나요?"

Rule 5: Realistic Search Behavior
The query must read as if a **researcher who does NOT have the document** is searching a database by topic and keywords. This single rule covers three aspects:

(a) No Verbatim Document Data
* Use **conceptual references** (policy names, years, entity names) instead of **exact figures**.
* **Strictly Forbidden**: Specific monetary values, exact percentages, precise statistics copied from the document.
  * Bad: "에너지 요금이 €49.5/MWh에서 €94/MWh로 89% 상승한 이유는?"
  * Good: "2022년 프랑스 소매 에너지 요금 급등과 EDF의 ARENH 정책은 어떤 관계가 있나요?"

(b) No Document-Aware Framing
* **Strictly Forbidden**: "문서에서", "해당 자료의", "위 표에 따르면", "본 보고서의", "제시된 데이터를 기반으로"
* Also forbidden — **Document Title Scoping** (assumes the user already knows the document exists):
  * Bad: "제7차 에너지기본계획에서 원전 비중 목표는?"
  * Good: "2025년 일본의 원전 비중 목표"

(c) Realistic User Knowledge
* The user **knows**: topic area, key entities, time periods of interest.
* The user **does NOT know**: page numbers, document structure, specific numerical values, exact document titles.

4. Query Format Specification

Question Format (질문형)
* Must be a complete interrogative sentence with question endings.
* **Required elements**: Question word (무엇, 어떻게, 왜, 어떤) OR question ending (~인가요?, ~있나요?, ~했는가?)
* **Examples**:
  * "M2 광의통화 증가율이 2020년 국가채무 증가에 영향을 미쳤는가?"
  * "에너지바우처 제도의 지원 대상은 누구인가요?"

Instruction Format (지시형)
* Must be a command with imperative endings.
* **Required elements**: Imperative ending (~해주세요, ~하시오, ~분석하라, ~설명하라)
* **Examples**:
  * "2020년 M2 통화량과 국가채무 간의 상관관계를 분석해주세요."
  * "에너지바우처와 에너지효율개선 사업의 차이점을 비교하라."

Keyword Format (키워드형)
* **NO complete sentences.** Only noun phrases and search terms.
* **NO verbs, NO question words, NO sentence endings.**
* Mimics search engine input: fragmented, noun-centric.

Keyword Format Rules:
Allowed / Forbidden
- 명사, 명사구, 복합 명사구 / 동사 (~하다, ~이다, ~있다)
- 관계 조사 (~의, ~와/과, ~간, ~에 따른, ~으로 인한) / 질문사 (무엇, 어떻게, 왜, 어떤)
- 고유명사, 연도, 날짜 / 문장 종결 (~인가요, ~해주세요, ~입니까)
- 관계 표현 (비교, 관계, 영향, 연관성, 상관관계) / 완전한 문장 구조
- 영문 약어 (EDF, ARENH, GDP) / 공백으로만 나열된 독립 키워드들
- 개념적 추상화 표현 / 문서 표 항목명·인덱스의 직접 복사

Keyword Structural Templates
A keyword query must form **one coherent noun phrase**. Every noun must be connected to its neighbors by Korean particles (의, 와/과, 간, 에 따른, 으로 인한, 내, 중, 및) that make the semantic relationship explicit.

Templates:
- Comparison: A의 X와/과 B의 Y (간) 차이/비교
  Example: "운수업의 부가가치당 에너지소비량과 수송용 에너지소비 비중 차이"
- Correlation: A와/과 B 간 연관성/관계/상관관계
  Example: "일반가구의 설계가중치와 도시가구의 에너지소비 간 연관성"

```
- Causation: A 변화/증가/감소와 B 변화의 연관성/영향
  Example: "부산 개별여행 비중 증가와 농수산물 구매 비중 상승의 연관성"
- Condition: A에 따른/으로 인한 B의 변화/추이
  Example: "스페인 용량요금 중단에 따른 전력부문 적자의 변화"
- Composition: A 내 B와 C의 비중/분포/구성
  Example: "EU 노동 인력 내 녹색 직업과 고도 디지털 집약 직업 간의 연령 분포"

**Particle Removal Test**:
Strip all particles (의/와/과/간/에 따른/으로 인한/내/중/및) from the query.
* If the meaning **collapses** -> Well-formed noun phrase.
* If the meaning **stays the same** -> Keyword bag. Rewrite.

**Read-Aloud Test**:
Read the query aloud. If there is a natural pause splitting it into two independent chunks
with no grammatical bridge -> Two queries glued together. Rewrite.

**Bad -> Fixed Examples**:
* "감일도서관 개관 희망도서 바로대출 지역서점 연계 독서문화 활성화 지원 사업 이동도서관
스마트도서관"
  -> "감일도서관 개관 이후 희망도서 바로대출 서비스와 지역서점 연계 독서문화 사업 간의 운영 방식
차이"
* "K-방산 폴란드 수출 비중 라틴아메리카 방위비 증가"
  -> "K-방산의 폴란드 수출 비중 확대와 라틴아메리카 방위비 증가 간 연관성"
* "베트남 최종 법인세 신고 베트남 개인소득세 체계 동일 과세 기준 여부"
  -> "베트남 법인세 최종 신고 체계와 개인소득세 체계의 과세 기준 동일 여부"

### 5. Quality Checklist (Self-Verification)
Before finalizing, verify ALL checks pass:
- Format Compliance: Query strictly follows the specified format
(question/instruction/keyword)
- Single Unit: ONE question, ONE instruction, or ONE keyword phrase — no multiple sentences
- No Page References: No page numbers, document indices, or positional references
- Realistic Search: No exact values from the document, no document-aware framing, no document
title scoping (Rule 5)
- Entity-Grounded: Includes searchable entities (names, years, policy names) but not verbatim
data
- Multi-Page Implicit: Requires information from multiple pages without explicitly stating it
- Keyword Coherence (keyword only): **Particle Removal Test** passes: stripping particles
must break the meaning
- Single Phrase (keyword only): **Read-Aloud Test** passes: query flows as one utterance with
no independent chunks

### 6. Output Generation
Generate the output strictly adhering to the defined JSON schema.
The query must be in **Korean** and must pass all checks in the Quality Checklist above.
**Pay special attention to the Query Format specification—the linguistic structure must
match exactly.**
```

Figure 11: Prompt for generating summary-based retrieval queries with controls for realism, diversity, and query formulation

```
You are an expert in creating challenging datasets for Vision Document Retrieval (VDR).
Your goal is to generate a **highly specific Korean search query** that acts as a realistic
user prompt for retrieving information from a large corpus.

### 1. Document Context
The following XML-like tags contain the markdown text extracted from a sequence of document
pages.
The <document index="..."> tags are for your internal reasoning ONLY. **Do NOT mention these
indices in the final query.**

<documents>
{% for doc in markdown %}
<document index="{{ loop.index0 }}">
{{ doc }}
</document>
{% endfor %}
</documents>

### 2. Task Requirements
You must generate a structured output containing the rationale and the query itself based on
the following specifications:

* **Query Type**: {{ query_type }} ({{ query_type_definition }})
* **Query Format**: {{ query_format }} ({{ query_format_definition }})

### 3. Critical Constraints for Realistic Retrieval

#### Rule 1: NO Artificial Location References
* **Strictly Forbidden**: "2페이지에서...", "다음 장에 있는...", "첫 번째 문서의...", "위에서
언급된..."
* **Reason**: The user queries the entire database and does not know the document order or
page numbers.
* **Alternative**: Use **Section Headers, Table Captions, or Unique Keywords** found in the
text.
  * Bad: "2페이지에 있는 표를 요약해."
  * Good: "'2024년 재무 하이라이트' 표를 요약해."

#### Rule 2: Implicit Multi-Page Synthesis
* The query must require information scattered across multiple pages, but **without
explicitly stating so**.
* **Strategy**: Identify **Entity A** on one page and **Entity B** on another, then ask about
their relationship.

#### Rule 3: Entity-Grounded Specificity
* Avoid generic queries like "에너지 정책을 분석해줘."
* Include specific entities found in the text: **Dates, Company Names, Regulations (e.g.,
ISO-27001), Project Codes, Policy Names, or Program Names**.
* However, do NOT include exact numerical values (see Rule 5).

#### Rule 4: Single Natural Query
* The query **MUST be a single unit** appropriate to its format (one question, one
instruction, or one keyword cluster).
* **Strictly Forbidden Patterns**:
  * Multiple sentences: "~입니다. ~해주세요."
  * Instruction suffixes: "단, ~를 기준으로 답변하시오."
  * Explicit output format requests: "~를 근거로 제시하시오.", "~를 나열하시오."
  * Conditional clauses at the end: "단, ~를 구분하여 제공해야 합니다."
* **Examples**:
  * Bad: "2021년과 2022년 상승률을 비교하시오. 단, 수도권과 지방을 구분하여 제시하시오."
  * Good: "2021년과 2022년 수도권 및 지방의 주택 매매가격 상승률은 어떻게 달랐나요?"

#### Rule 5: Realistic Search Behavior
The query must read as if a **researcher who does NOT have the document** is searching a
database by topic and keywords. This single rule covers three aspects:

**(a) No Verbatim Document Data**
* Use **conceptual references** (policy names, years, entity names) instead of **exact
figures**.
* **Strictly Forbidden**: Specific monetary values, exact percentages, precise statistics
copied from the document.
  * Bad: "에너지 요금이 €49.5/MWh에서 €94/MWh로 89% 상승한 이유는?"
  * Good: "2022년 프랑스 소매 에너지 요금 급등과 EDF의 ARENH 정책은 어떤 관계가 있나요?"
```

(b) No Document-Aware Framing
* **Strictly Forbidden**: "문서에서", "해당 자료의", "위 표에 따르면", "본 보고서의", "제시된 데이터를 기반으로"
* Also forbidden — **Document Title Scoping** (assumes the user already knows the document exists):
  * Bad: "제7차 에너지기본계획에서 원전 비중 목표는?"
  * Good: "2025년 일본의 원전 비중 목표"

(c) Realistic User Knowledge
* The user **knows**: topic area, key entities, time periods of interest.
* The user **does NOT know**: page numbers, document structure, specific numerical values, exact document titles.

4. Query Format Specification

Question Format (질문형)
* Must be a complete interrogative sentence with question endings.
* **Required elements**: Question word (무엇, 어떻게, 왜, 어떤) OR question ending (~인가요?, ~있나요?, ~했는가?)
* **Examples**:
  * "M2 광의통화 증가율이 2020년 국가채무 증가에 영향을 미쳤는가?"
  * "에너지바우처 제도의 지원 대상은 누구인가요?"

Instruction Format (지시형)
* Must be a command with imperative endings.
* **Required elements**: Imperative ending (~해주세요, ~하시오, ~분석하라, ~설명하라)
* **Examples**:
  * "2020년 M2 통화량과 국가채무 간의 상관관계를 분석해주세요."
  * "에너지바우처와 에너지효율개선 사업의 차이점을 비교하라."

Keyword Format (키워드형)
* **NO complete sentences.** Only noun phrases and search terms.
* **NO verbs, NO question words, NO sentence endings.**
* Mimics search engine input: fragmented, noun-centric.

Keyword Format Rules:
Allowed / Forbidden
- 명사, 명사구, 복합 명사구 / 동사 (~하다, ~이다, ~있다)
- 관계 조사 (~의, ~와/과, ~간, ~에 따른, ~으로 인한) / 질문사 (무엇, 어떻게, 왜, 어떤)
- 고유명사, 연도, 날짜 / 문장 종결 (~인가요, ~해주세요, ~입니까)
- 관계 표현 (비교, 관계, 영향, 연관성, 상관관계) / 완전한 문장 구조
- 영문 약어 (EDF, ARENH, GDP) / 공백으로만 나열된 독립 키워드들
- 개념적 추상화 표현 / 문서 표 항목명·인덱스의 직접 복사

Keyword Structural Templates
A keyword query must form **one coherent noun phrase**. Every noun must be connected to its neighbors by Korean particles (의, 와/과, 간, 에 따른, 으로 인한, 내, 중, 및) that make the semantic relationship explicit.

Templates:
- Comparison: A의 X와/과 B의 Y (간) 차이/비교
  Example: "운수업의 부가가치당 에너지소비량과 수송용 에너지소비 비중 차이"
- Correlation: A와/과 B 간 연관성/관계/상관관계
  Example: "일반가구의 설계가중치와 도시가구의 에너지소비 간 연관성"
- Causation: A 변화/증가/감소와 B 변화의 연관성/영향
  Example: "부산 개별여행 비중 증가와 농수산물 구매 비중 상승의 연관성"
- Condition: A에 따른/으로 인한 B의 변화/추이
  Example: "스페인 용량요금 중단에 따른 전력부문 적자의 변화"
- Composition: A 내 B와 C의 비중/분포/구성
  Example: "EU 노동 인력 내 녹색 직업과 고도 디지털 집약 직업 간의 연령 분포"

Particle Removal Test:
Strip all particles (의/와/과/간/에 따른/으로 인한/내/중/및) from the query.
* If the meaning **collapses** -> Well-formed noun phrase.
* If the meaning **stays the same** -> Keyword bag. Rewrite.

Read-Aloud Test:
Read the query aloud. If there is a natural pause splitting it into two independent chunks with no grammatical bridge -> Two queries glued together. Rewrite.

Bad -> Fixed Examples:
* "감일도서관 개관 희망도서 바로대출 지역서점 연계 독서문화 활성화 지원 사업 이동도서관 스마트도서관"

```
  -> "감일도서관 개관 이후 희망도서 바로대출 서비스와 지역서점 연계 독서문화 사업 간의 운영 방식
차이"
* "K-방산 폴란드 수출 비중 라틴아메리카 방위비 증가"
  -> "K-방산의 폴란드 수출 비중 확대와 라틴아메리카 방위비 증가 간 연관성"
* "베트남 최종 법인세 신고 베트남 개인소득세 체계 동일 과세 기준 여부"
  -> "베트남 법인세 최종 신고 체계와 개인소득세 체계의 과세 기준 동일 여부"

### 5. Quality Checklist (Self-Verification)
Before finalizing, verify ALL checks pass:
- Format Compliance: Query strictly follows the specified format
(question/instruction/keyword)
- Single Unit: ONE question, ONE instruction, or ONE keyword phrase - no multiple sentences
- No Page References: No page numbers, document indices, or positional references
- Realistic Search: No exact values from the document, no document-aware framing, no document
title scoping (Rule 5)
- Entity-Grounded: Includes searchable entities (names, years, policy names) but not verbatim
data
- Multi-Page Implicit: Requires information from multiple pages without explicitly stating it
- Keyword Coherence (keyword only): **Particle Removal Test** passes: stripping particles
must break the meaning
- Single Phrase (keyword only): **Read-Aloud Test** passes: query flows as one utterance with
no independent chunks

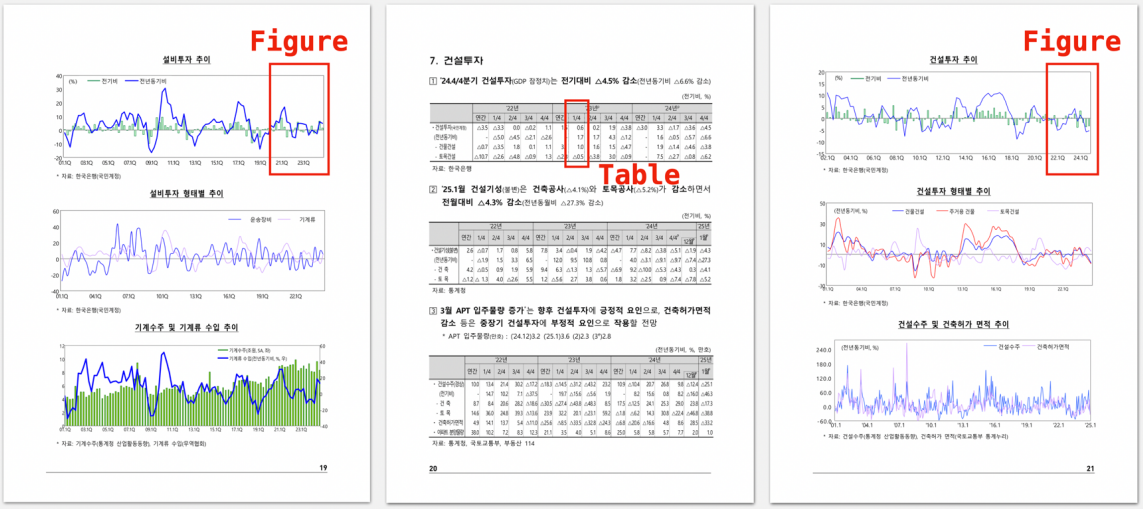
### 6. Output Generation
Generate the output strictly adhering to the defined JSON schema.
The query must be in **Korean** and must pass all checks in the Quality Checklist above.
**Pay special attention to the Query Format specification—the linguistic structure must
match exactly.**
```

Figure 12: Prompt for generating context-based retrieval queries with controls for realism, diversity, and query formulation

Economic

🔎 **Query:** 설비투자와 건설 투자의 전기비 및 전년동기비 추이는 2023년 1분기까지 어떻게 달랐나요?

Relevant Pages



Energy

🔎 **Query:** 에너지사용량 신고제도와 에너지진단 제도 대상 절차 비교

Relevant Pages

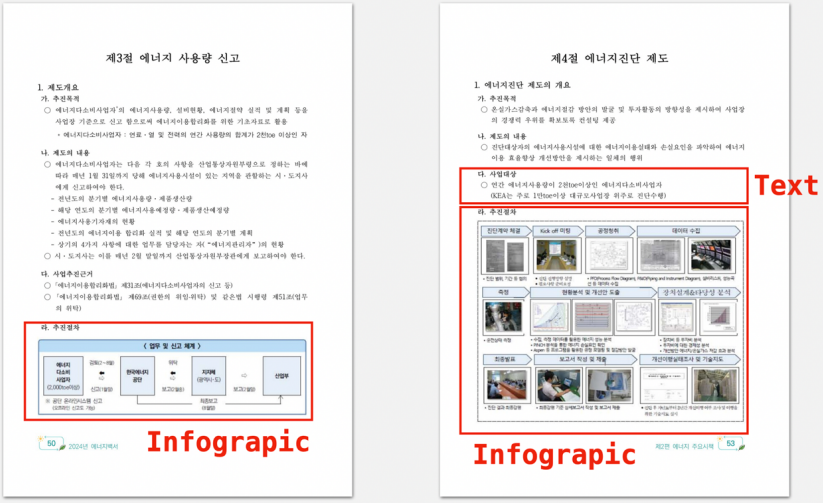
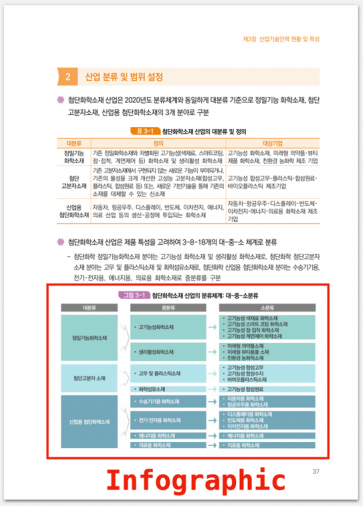


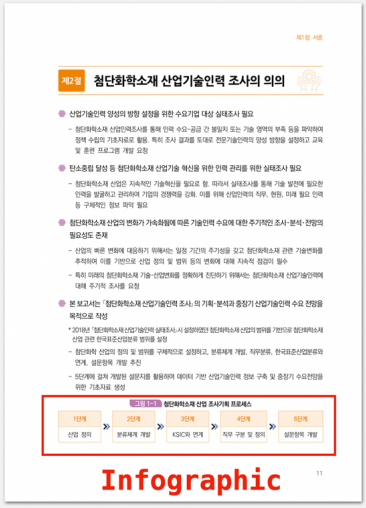
Figure 13: Example query-page pairs from the Economic and Energy subsets. Highlighted regions indicate the key evidence supporting each query.

HR

🔎 **Query:** 첨단화학소재의 5단계 조사 프로세스와 산업 분류 체계의 계층 구조를 분석하라.

Relevant Pages



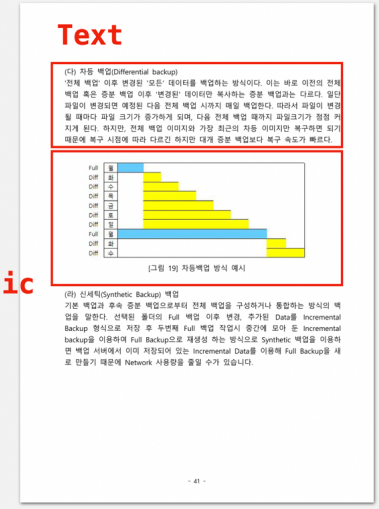


Cybersecurity

🔎 **Query:** 차등 백업과 중복제거 백업 방식의 저장 효율성과 복구 과정에서 어떤 차이가 있나요?

Relevant Pages

Text



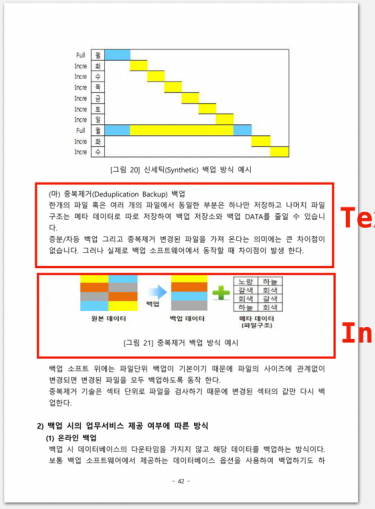
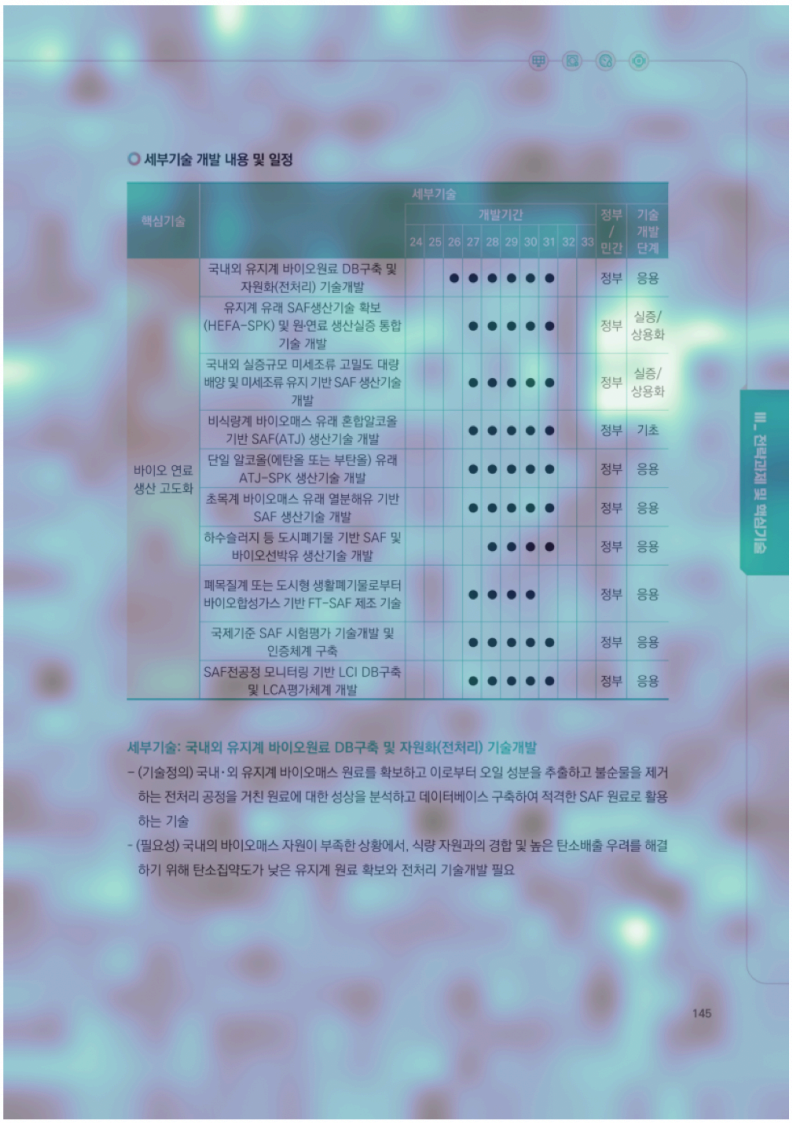


Figure 14: Example query-page pairs from the HR and Cybersecurity subsets. Highlighted regions indicate the key evidence supporting each query.



Query: **실증/상용화** 단계에 있는 SAF 생산기술의 총 개수는 몇 개인가요?

Figure 15: Additional query-to-document similarity heatmap for the fine-tuned `colqwen2-v1.0` model on a Korean document example. The model assigns high similarity to image patches corresponding to the term "실증/상용화," highlighting its focus on query-relevant textual evidence.

Title	Provider	Pages	License
Cybersecurity			
갠드크랩 랜섬웨어 악성코드 분석 기술 보고서	한국인터넷진흥원	92	No Restriction
사이버위협 동향보고서 (Windows 취약점 동향 및 업데이트 정책 등)	한국인터넷진흥원	104	No Restriction
사이버위협 동향보고서 (동형암호 기반 데이터 결합 및 분석 등)	한국인터넷진흥원	100	No Restriction
사이버 위협 동향보고서 (피싱 메일 공격 사례 등)	한국인터넷진흥원	96	No Restriction
사이버 위협 동향보고서 (기업 보안관리자의 크리덴셜 스터핑(Credential Stuffing) 공격 대응방안 등)	한국인터넷진흥원	104	No Restriction
사이버위협 동향보고서 (공인인증서 문제점과 DID 기술 전망 등)	한국인터넷진흥원	100	No Restriction
사이버위협 동향보고서 (ATT&CK Framework 개념과 이해 등)	한국인터넷진흥원	80	No Restriction
랜섬웨어 대응을 위한 안전한 정보시스템 백업 가이드(개정본)	한국인터넷진흥원	68	No Restriction
해킹진단도구 활용 사례 (취약한 관리자 계정 악용을 악용한 데이터 유출)	한국인터넷진흥원	23	No Restriction
해킹진단도구 활용 사례 (노출된 SMB 파일 서버를 통한 AD 환경 장악)	한국인터넷진흥원	23	No Restriction
해킹진단도구 활용 사례 (취약한 MS-SQL 서버를 통한 랜섬웨어 침해사고)	한국인터넷진흥원	14	No Restriction
해킹진단도구 활용 사례 (AD 환경에서의 RAT 악성코드 감염)	한국인터넷진흥원	16	No Restriction
AD서버 악용 내부망 랜섬웨어 유포 사례 분석	한국인터넷진흥원	31	No Restriction
Log4j 위협 대응 보고서	한국인터넷진흥원	35	No Restriction
NAS 보안 가이드	한국인터넷진흥원	161	No Restriction
TTPs2 스피어 피싱을 통한 공격망 구성 방식 분석	한국인터넷진흥원	79	No Restriction
TTPs3 공격자의 악성코드 활용 전략 분석	한국인터넷진흥원	27	No Restriction
Economic			
최근 경제동향 (2021. 3월호)	기획재정부	80	No Restriction
최근 경제동향 (2021. 6월호)	기획재정부	80	No Restriction
최근 경제동향 (2021. 9월호)	기획재정부	80	No Restriction
최근 경제동향 (2021. 12월호)	기획재정부	80	No Restriction
최근 경제동향 (2022. 3월호)	기획재정부	80	No Restriction
최근 경제동향 (2022. 6월호)	기획재정부	80	No Restriction
최근 경제동향 (2022. 9월호)	기획재정부	80	No Restriction
최근 경제동향 (2022. 12월호)	기획재정부	80	No Restriction
최근 경제동향 (2023. 3월호)	기획재정부	82	No Restriction
최근 경제동향 (2023. 6월호)	기획재정부	80	No Restriction
최근 경제동향 (2023. 9월호)	기획재정부	80	No Restriction
최근 경제동향 (2023. 12월호)	기획재정부	80	No Restriction
최근 경제동향 (2024. 3월호)	기획재정부	77	No Restriction
최근 경제동향 (2024. 6월호)	기획재정부	77	No Restriction
최근 경제동향 (2024. 9월호)	기획재정부	77	No Restriction
최근 경제동향 (2024. 12월호)	기획재정부	77	No Restriction
최근 경제동향 (2025. 3월호)	기획재정부	77	No Restriction
최근 경제동향 (2025. 6월호)	기획재정부	77	No Restriction
최근 경제동향 (2025. 9월호)	기획재정부	77	No Restriction
최근 경제동향 (2025. 12월호)	기획재정부	77	No Restriction
Energy			
해외전력산업동향 (2017 China)	한국전력거래소	57	No Restriction
해외전력산업동향 (2017 Japan)	한국전력거래소	43	No Restriction
해외전력산업동향 (2017 USA)	한국전력거래소	43	No Restriction
제4차 에너지기술개발계획 기술로드맵: 에너지저장	한국에너지기술평가원	172	KOGL Type 2
제4차 에너지기술개발계획 기술로드맵: 총괄	한국에너지기술평가원	76	KOGL Type 2
대전광역시 신재생에너지 보급계획	대전광역시	536	No Restriction
해외전력산업동향 (2023)	한국전력거래소	478	No Restriction
인천광역시 에너지백서	인천광역시	259	No Restriction
제5차 에너지기술개발계획 기술로드맵: 수요관리	한국에너지기술평가원	85	KOGL Type 2
제5차 에너지기술개발계획 기술로드맵: 효율향상	한국에너지기술평가원	162	KOGL Type 2
에너지 기술정책 포커스 (2025 주요국 기후에너지정책)	한국에너지기술연구원	122	No Restriction
HR			
고용형태별 근로실태조사 보고서	고용노동부	277	KOGL Type 1
블라인드 채용 가이드북	고용노동부	88	No Restriction
일 · 가정 양립 실태조사 보고서	고용노동부	423	No Restriction
한국직업전망	한국고용정보원	668	KOGL Type 2
유망 신산업 산업기술인력 전망: 이차전지	한국산업기술진흥원	134	No Restriction
유망 신산업 산업기술인력 전망: 첨단화학소재	한국산업기술진흥원	134	No Restriction
유망 신산업 산업기술인력 전망: 첨단섬유소재	한국산업기술진흥원	148	No Restriction
유망 신산업 산업기술인력 전망: 신금속소재	한국산업기술진흥원	136	No Restriction
유망 신산업 산업기술인력 전망: 차세대세라믹소재	한국산업기술진흥원	138	No Restriction

Table 7: Document metadata for KoViDoRe across all subsets.

Title	Provider	Pages	License
에너지총조사	기후에너지환경부	774	No Restrictions
주요업무계획	경기도 하남시	640	No Restrictions
지방공무원 인사실무	행정안전부	521	No Restrictions
항만편람	해양수산부	515	No Restrictions
국가연구개발사업 상위평가보고서	과학기술정보통신부	479	No Restrictions
연구보고서 현황	한국방송통신전파진흥원	425	No Restrictions
작업환경실태조사 보고서	한국산업안전보건공단	363	No Restrictions
국가연구개발사업 특정평가보고서	과학기술정보통신부	323	No Restrictions
관리형매립지 조사결과보고서	수도권매립지관리공사	285	No Restrictions
ICT 융복합 시설의 안전한 전자파 환경 기반 조성 연구	국립전파연구원	253	KOGL Type 1
전자파 흡수전력밀도 등 전자파 인체노출량 평가기술 연구	국립전파연구원	249	KOGL Type 1
해외건설 세무업무 매뉴얼	국토교통부	216	No Restrictions
처분시설 부지주변 방사선환경조사 보고서	한국원자력환경공단	211	KOGL Type 1
스마트 안전유지관리 시설물 확대방안 마련 용역 보고서	국토안전관리원	210	No Restrictions
해양수산발전기본계획	해양수산부	204	No Restrictions
디지털미디어 허브 조성을 위한 빛마루 중장기 전략 연구보고서	한국방송통신전파진흥원	195	No Restrictions
유엔개황 책자	외교부	178	No Restrictions
기업체노동비용조사 보고서	고용노동부	153	No Restrictions
(PDF)인삼재배전서	경상북도	151	No Restrictions
디지털미디어 신산업 진흥 방안 및 인력수급 기초조사에 관한 연구보고서	한국방송통신전파진흥원	146	No Restrictions
무인도서 100선	해양수산부	125	KOGL Type 1
환경관리해역 기본계획	해양수산부	125	No Restrictions
해외건설 법률컨설팅 사례	국토교통부	121	No Restrictions
국내외 온라인 동영상 미디어 콘텐츠 시장 전망 및 정책 추진방향 연구보고서	한국방송통신전파진흥원	115	No Restrictions
합성데이터 생성 활용 안내서	개인정보보호위원회	110	KOGL Type 1
환경관리해역 기본계획	해양수산부	90	No Restrictions
국가공무원인재개발원 교육운영계획	인사혁신처	85	No Restrictions
중소기업 경제동향 정보	중소벤처기업연구원	75	No Restrictions
생체정보 보호 안내서	개인정보보호위원회	73	KOGL Type 1
인재개발 종합계획	인사혁신처	68	No Restrictions
공공외교 기본계획	외교부	59	No Restrictions
관광실태조사 정보	부산광역시	58	No Restrictions
카지노 비즈니스와 제도	그랜드코리아레저(주)	57	No Restrictions
모바일 전자정부서비스 앱 소스코드 검증 가이드라인	행정안전부	51	No Restrictions
개인정보 유출 등 사고 대응 매뉴얼	개인정보보호위원회	48	KOGL Type 1
교통 리포트	서울특별시	42	No Restrictions
블랙잭 게임의 이해	그랜드코리아레저(주)	32	No Restrictions
개인정보 유출 신고 동향 및 예방 방법	한국인터넷진흥원	32	No Restrictions
i SMR 및 SSNC 설명자료	한국수력원자력(주)	27	No Restrictions
농지개량행위 신고 업무지침	농림축산식품부	24	No Restrictions
미디어이슈_광고요금제 도입을 앞둔 넷플릭스에 대한 인식 및 이용 조사	한국언론진흥재단	23	KOGL Type 1
위성전파 감시 정보	중앙전파관리소	19	KOGL Type 1
미디어이슈_이대남 현상에 대한 인식	한국언론진흥재단	19	KOGL Type 1
미디어이슈_코로나19 관련 정보 이용 및 인식 현황	한국언론진흥재단	19	KOGL Type 1
해외시장 신용위험 보고서	한국무역보험공사	18	No Restrictions
중남미 관련 보고서 (제약바이오)	외교부	12	No Restrictions
중남미 관련 보고서 (방위산업)	외교부	10	No Restrictions
노인 일자리 및 사회활동 지원사업 시행 20년의 성과와 발전과제	한국노인인력개발원	9	No Restrictions
전국 주배관망 가스인입지점별 인입가능량	한국가스공사	3	No Restrictions

Table 8: Document metadata for Ko-VDR Train Public.

Decoupling Semantics and Logic: A Training-Free Coarse-to-Fine Pipeline for Video Retrieval-Augmented Generation

Jiaxin Dai[2*], **Zehang Wei**[2*], **Jiamin Yan**[2*], **Xiang Xiang**[1*]
[1]School of Computer Science & Tech, Huazhong University of Science and Technology
[2]School of AI and Automation, Huazhong University of Science and Technology, China
xex@hust.edu.cn

Abstract

This paper presents our system description for the 2nd Workshop on Multimodal Augmented Generation via MultimodAl Retrieval (MAGMaR). Addressing the critical challenges of cross-lingual long-video comprehension, strict persona adherence, and zero-hallucination temporal grounding, we propose a fully training-free, two-stage cascaded Video RAG pipeline. Our architecture strategically decouples semantic retrieval from cognitive logical reasoning through a modality-aware division of labor. In the first stage, a high-recall semantic pre-fetching module employs dense retrieval using only high-fidelity visual summaries and global text descriptions, explicitly isolating noisy modalities (e.g., OCR and ASR) to maintain a pristine vector space. In the second stage, an Adaptive, Iterative, and Reasoning-based (A.I.R.) filtering agent, powered by a commercial Large Language Model (LLM), performs fine-grained cognitive reranking. The agent re-incorporates full multimodal contexts to enforce strict logical alignment with user personas, effectively pruning semantically similar but logically irrelevant candidates. Finally, a Prompt Sculpting mechanism constrains the generator to synthesize the distilled subset into strictly formatted JSON responses with exact chunk-level citations. Evaluated on the RAG track, our resource-aware approach shows exceptional precision in both information retrieval and persona-conditioned generation.

1 Introduction

The rapid proliferation of multimodal content has intensified the demand for systems capable of synthesizing complex information from extensive video archives. The 2nd Workshop on MAGMaR addresses this challenge by requiring systems to generate persona-constrained responses grounded in multiple retrieved videos. This paradigm builds upon the foundational challenge of synthesizing coherent narratives from heterogeneous video sources, as pioneered by benchmarks such as WikiVideo (Martin et al., 2025a).

Unlike text-based Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), Video RAG must navigate the high-dimensional noise of long-form video, cross-lingual barriers, and the requirement for precise temporal grounding across massive collections (Lei et al., 2021; Gao et al., 2017).

A fundamental bottleneck in current Video RAG systems is the heavy reliance on surface-level semantic similarity for retrieval. While dense vector embeddings effectively capture visual and textual correlations, they often struggle to distinguish between a "semantically similar" distractor and a "logically relevant" evidence. This "semantic-logical gap" is particularly pronounced when a system is tasked with following a specific persona, where the relevance of a video is determined not just by its visual content, but by its logical alignment with a particular viewpoint or query nuance. Single-stage retrieval often leads to high recall but low precision, introducing "hard negatives" (Xiong et al., 2020) that trigger downstream hallucinations (Ji et al., 2023) in the generation phase.

To address these challenges, we propose **C2F-RAG** (Coarse-to-Fine RAG), a fully training-free, two-stage cascaded pipeline designed to decouple semantic fetching from cognitive logical reasoning. Our approach is grounded in a coarse-to-fine philosophy. In the first stage, we prioritize high-recall semantic pre-fetching by leveraging BGE-M3, a state-of-the-art embedding model capable of multi-lingual and multi-granularity representation. By utilizing lightweight, global visual and textual summaries, BGE-M3 efficiently embeds the vast corpus into a dense vector space, enabling rapid candidate retrieval while filtering out the vast majority of irrelevant background noise without prohibitive computational overhead.

*Euqal contribution, co-first author.

Proceedings of the 2nd Workshop on Multimodal Augmented Generation via Multimodal Retrieval, pages 81–91

In the second stage, we adapt the Adaptive, Iterative, and Reasoning-based (A.I.R.) framework (Zou et al., 2025) to design a multimodal cognitive filtering agent. While the original A.I.R. focuses on optimizing intra-video frame selection for VideoQA, our tailored agent elevates this mechanism to tackle inter-video cognitive reranking in large-scale Video RAG. Utilizing the advanced capabilities of a commercial Large Language Model (LLM), our agent processes the comprehensive multimodal context of each retrieved candidate (including OCR and ASR data) and deeply evaluates them against strict persona constraints. This allows the system to effectively prune logically irrelevant videos that surface-level embeddings fail to filter. The cascaded architecture ensures that the final generator operates exclusively on a distilled "golden subset," significantly reducing the risk of knowledge injection from irrelevant distractors.

Our primary contributions are as follows:

- We propose C2F-RAG, a two-stage cascaded pipeline that explicitly decouples semantic retrieval (via BGE-M3) from logical reasoning (via an adapted A.I.R. agent), bridging the semantic-logical gap in large-scale retrieval.
- We present a tailored cognitive reranking strategy that leverages LLM-driven logical alignment and persona constraints to effectively prune hard negatives and prevent downstream hallucinations.
- We show a training-free synthesis approach that utilizes prompt sculpting to enforce strict JSON formatting and accurate chunk-level temporal grounding, achieving state-of-the-art precision without prior fine-tuning.

2 System Architecture

In this section, we detail the design of C2F-RAG, a two-stage cascaded pipeline optimized for large-scale multimodal retrieval and persona-constrained generation. Our system adopts a "coarse-to-fine" philosophy (Karpukhin et al., 2020), strategically decoupling surface-level semantic fetching from deep cognitive logical reasoning to navigate the noise inherent in massive video collections.

2.1 Overview

The core of C2F-RAG is a cascaded data flow designed to distill a high-fidelity "golden subset" from a vast search space. As illustrated in Figure 1, the pipeline consists of two primary stages:

1. **Coarse Stage: High-Recall Semantic Prefetching.** This stage utilizes BGE-M3 for dense retrieval over lightweight global visual and textual summaries. It efficiently narrows down the 110k video corpus to a candidate pool of Top-1000 items.
2. **Fine Stage: Cognitive Reranking via Serial Multimodal Context.** In this stage, the system re-integrates fine-grained modalities (e.g., OCR and ASR) and serializes them into what we define as the Serial Multimodal Context (SMC). An adapted A.I.R. agent then performs logical alignment against the query and persona constraints, pruning hard negatives to derive the final context for generation.

By bridging these stages, C2F-RAG maintains a balance between computational scalability and reasoning depth, ensuring that the downstream generator receives only the most logically relevant and contextually rich evidence.

2.2 The Coarse Stage: High-Recall Semantic Pre-fetching

The primary objective of the Coarse Stage is to efficiently narrow the search space from the entire 110k video corpus to a manageable candidate pool of Top-1000 videos. Given the massive scale of the background collection, this stage prioritizes high recall and computational scalability while maintaining semantic understanding of the video content.

Modality Decoupling for High SNR To construct a pristine and efficient vector space, we implement a strategy of Modality Decoupling. During the indexing phase, we exclude noisy or fragmented modalities such as OCR and ASR. We observe that while these modalities provide fine-grained evidence, their high variance and localized nature introduce significant noise into dense vector representations during large-scale retrieval. Instead, we represent each video using a "Sandwich" text structure consisting of two distinct components:

- A high-level Global Summary generated by a multimodal LLM (Qwen-Omni) (Bai et al., 2023; Zhu et al., 2024).
- Dense Frame Details consisting of visual captions for keyframes.

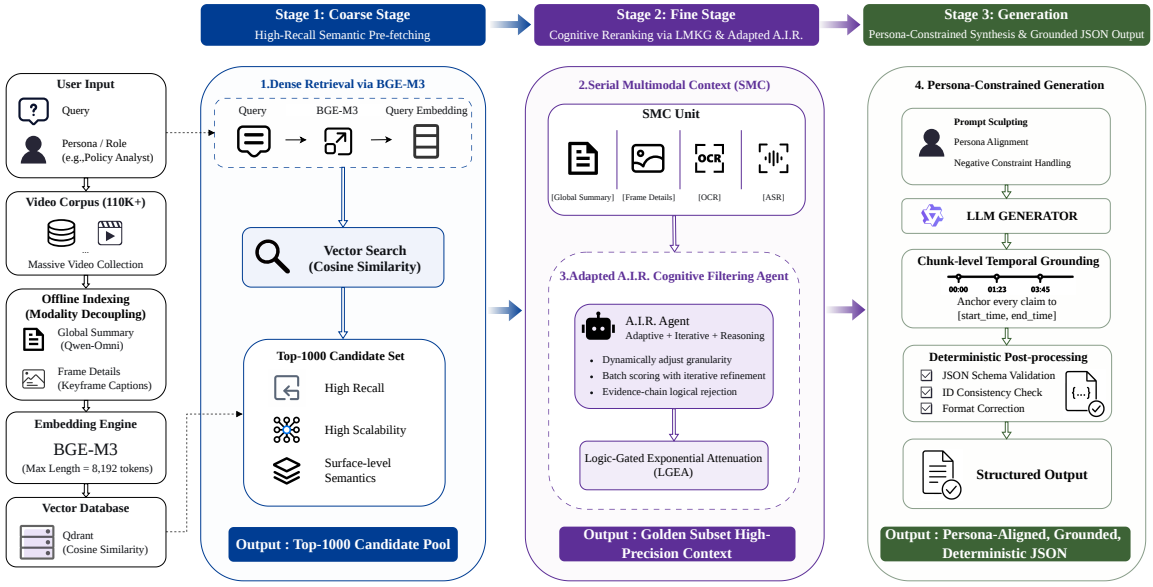


Figure 1: The overall architecture of the C2F-RAG pipeline. The system operates in three cascaded stages: (1) **Coarse Stage**: High-recall semantic pre-fetching utilizes BGE-M3 to retrieve a Top-1000 candidate pool from the 110K+ video corpus, based on decoupled global summaries and visual frames. (2) **Fine Stage**: Deep cognitive reranking re-integrates fine-grained modalities (OCR, ASR) into the Serial Multimodal Context (SMC). An adapted A.I.R. agent performs iterative, logic-gated filtering to distill a high-precision Golden Subset. (3) **Generation Stage**: Persona-constrained synthesis leverages the LLM to produce chunk-level temporally grounded responses.

This combination ensures a high signal-to-noise ratio (SNR) for initial semantic matching.

Dense Retrieval via BGE-M3 We utilize BGE-M3 (Chen et al., 2024) as our core embedding engine. BGE-M3 is chosen for its superior support for multi-lingual queries and its ability to handle extremely long sequences. Specifically, we set the maximum sequence length to 8,192 tokens to ensure that the comprehensive visual-textual descriptions of long-form videos are fully captured without truncation. The video embeddings are indexed in a high-performance vector database (*Qdrant*) (Malkov and Yashunin, 2018), using cosine similarity as the distance metric.

Candidate Pool Generation For each user query, the system performs a dense vector search to retrieve the Top-1000 candidates. This wide-recall strategy ensures that even if the ground-truth video is logically complex, it is captured within the candidate pool based on its surface-level semantic features. This pool serves as the raw input for the subsequent fine-grained logical filtering.

2.3 The Fine Stage: Cognitive Reranking via Adapted A.I.R.

The Fine Stage represents the transition from surface-level semantic matching to deep logical alignment. To bridge the "semantic-logical gap," we implement an adapted version of the A.I.R. framework, functioning as a cognitive filtering agent that operates over a specialized data structure we define as the Serial Multimodal Context.

Multimodal Context Serialization To provide the agent with high-fidelity evidence, we reconstruct the multimodal data for each candidate in the Top-1000 pool. We implement a Serial Multimodal Context (SMC) approach to integrate heterogeneous data. This mechanism flattens the decoupled video elements—including global summaries, localized visual frames, and time-aligned OCR/ASR transcripts—into a continuous, time-ordered textual representation using specific modality identifiers (e.g., `[Global Summary]`, `[OCR]`, `[ASR]`). By serializing these multimodal streams, we enable the downstream LLM-based reasoning agent to perform cross-modal evidence synthesis within a unified context window, preserving the chronological integrity of the original video source.

The A.I.R. Mechanism Our cognitive reranking is driven by three core principles derived from the A.I.R. philosophy:

- **Adaptive (A):** The system dynamically manages the context window by adapting the

granularity of the SMC based on the query complexity. For intricate persona constraints, the agent prioritizes high-entropy modalities (OCR/ASR) to ensure precise logical grounding, while for general queries, it relies more on visual-textual summaries to maintain computational efficiency.

- **Iterative (I):** Given the substantial data volume of the SMC for 1,000 candidates, processing all videos in a single pass is computationally prohibitive. We employ an Iterative Refinement Loop (Shinn et al., 2023; Asai et al., 2023). The candidate pool is divided into batches (e.g., 20 videos per reasoning unit). The agent iteratively evaluates each batch, incrementally populating and refining the "Golden Subset" of 10-15 videos. This iterative pruning ensures that the system remains scalable while maintaining exhaustive logical oversight.

- **Reasoning-based (R):** Unlike embedding models that rely on lexical overlap, our agent performs deep Cognitive Reasoning. It evaluates each video against the target persona and query using a "Chain-of-Evidence" prompt (Wang et al., 2023). If the SMC contains evidence that is semantically similar but logically contradicts the query (a "hard negative"), the agent explicitly rejects the candidate.

Logic-Gated Exponential Attenuation To effectively bridge the semantic-logical gap while strictly bounding the score distribution within $[0, 1]$, we implement a Logic-Gated Exponential Attenuation (LGEA) mechanism. Rather than artificially inflating scores with heuristic margins, the final relevance score S_{final} for a video candidate v is derived by preserving the base semantic score $s_{coarse}(v)$ and exponentially penalizing logical inconsistency evaluated by the A.I.R. agent:

$$S_{final}(v) = s_{coarse}(v) \cdot \exp\left(-\gamma \cdot (1 - L(v))\right)$$

where $s_{coarse} \in [0, 1]$ represents the initial dense retrieval similarity, $L(v) \in [0, 1]$ denotes the logical alignment confidence from the A.I.R. agent, and $\gamma > 0$ is the *attenuation hyperparameter*.

This formulation provides a mathematically rigorous "soft-hard margin." When a candidate perfectly aligns with the persona ($L = 1$), its semantic score remains perfectly preserved ($\exp(0) = 1$). When a candidate represents a semantic distractor ($L = 0$), its score is aggressively suppressed by a factor of $e^{-\gamma}$. Crucially, tuning γ allows the system to balance reasoning strictness with semantic fidelity. Under an optimal γ, a candidate with abysmal semantic similarity will not unilaterally bypass a highly relevant semantic match simply due to logical alignment, thereby maintaining the topological integrity of the original vector space.

2.4 Persona-Constrained Generation and Temporal Grounding

The final stage of C2F-RAG is responsible for synthesizing the distilled evidence into a persona-consistent response. To meet the demanding requirements of the MAGMaR task, we employ a strategy termed Prompt Sculpting to ensure strict adherence to complex user personas and deterministic formatting.

Prompt Sculpting for Logical Synthesis The generator receives the "Golden Subset" of videos (typically 10-15 candidates) identified by the Fine Stage, along with their associated SMC content. We define Prompt Sculpting (Lu et al., 2022) as the process of structuring the LLM's task through a multi-layered instructional template. This template explicitly enforces two critical constraints:

- **Persona Alignment:** The LLM is forced to adopt the specified role (e.g., policy analyst, eyewitness) and filter information through that cognitive lens.

- **Negative Constraint Handling:** Explicit instructions to ignore any candidates that do not strictly meet the evidence threshold, thus reinforcing the zero-hallucination objective.

Chunk-level Temporal Grounding A critical requirement for MAGMaR is the precise mapping of information to temporal segments. Unlike traditional RAG which cites entire documents, C2F-RAG performs Chunk-level Temporal Grounding. By utilizing the timestamped metadata preserved within the SMC (from OCR and ASR modalities), the generator is instructed to anchor every factual claim to a specific time interval ($[start\_time, end\_time]$). This fine-grained citation ensures the generated output is fully auditable and rooted in verifiable visual or auditory evidence.

Deterministic Schema Enforcement To guarantee a 100% compliance rate with the submission

requirements, the system utilizes a constrained decoding approach (Willard and Louf, 2023; Scholak et al., 2021). We define a strict JSON schema that the LLM must follow, including mandatory fields for query IDs, video IDs, and the synthesized response. Any minor structural deviations in the raw LLM output are corrected via a Deterministic Post-processing layer, which validates the JSON integrity and ensures that all cited video IDs correspond to the actual retrieved candidates. This architectural safeguard eliminates common formatting errors and ensures that the system's high-precision reasoning is perfectly serialized for evaluation.

3 Experiments

To demonstrate the robustness and precision of C2F-RAG, we evaluate the system on the MAGMaR Full RAG track. Our experiments are designed to answer two core questions:

1. Can the two-stage cognitive filtering architecture effectively isolate true evidence from massive background noise compared to standard retrieval baselines?

2. Does the system maintain high-fidelity persona constraints and temporal grounding when scaling to real-world corpus sizes?

3.1 Experimental Setup

Dataset We evaluate on the MAGMaR2026 Test Set. The data is based on WikiVideo (Martin et al., 2025a). For the retrieval and RAG settings, we retrieve relevant videos from a combination of the MAGMaR data and MultiVENT2.0 test (Kriz et al., 2025). The background collection comprises $\sim 110,000$ multilingual, event-centric videos.

Retrieval Evaluation Setup Retrieval results are evaluated with nDCG and Recall for 10, 20, and 100. We use ir-measures (MacAvaney et al., 2022) to calculate these scores.

Generation Evaluation Setup Predictions are evaluated using an automatic evaluation framework. Specifically, the systems are evaluated by MiRAGE (Martin et al., 2025b), which captures the factuality, information coverage, groundedness, and proper attribution of citations. Each MiRAGE entailment judgment is judged by CLUE (Zhang et al., 2026).

Baselines To benchmark the effectiveness of our retrieval pipeline, we compare C2F-RAG against several official and industry-standard baselines provided in the MAGMaR leaderboard:

- **OmniEmbed** (Ma et al., 2025): A foundational zero-shot multimodal embedding baseline that relies purely on single-stage dense vector similarity.

- **OmniEmbed + RankVideo** (Skow et al., 2026): A traditional two-stage pipeline that employs OmniEmbed for initial recall and a standard RankVideo module for re-ranking.

- **Mixedbread** (Lee et al., 2024; Li and Li, 2023): A highly competitive, commercial state-of-the-art dense retrieval system known for its robust semantic matching capabilities.

3.2 Retrieval Performance: The Power of Cognitive Filtering

The primary bottleneck in scaling Video RAG to 110,000 videos is the degradation of ranking quality due to the intrusion of "hard negatives"—videos that share superficial semantic similarities but lack logical alignment with the query and persona. Table 1 summarizes the comparative retrieval performance against official baselines.

Overcoming the Semantic-Logical Gap As observed in Table 1, traditional single-stage semantic retrieval (OmniEmbed) collapses under the noise of the 110k corpus, achieving an nDCG@10 of only 0.166. While adding a standard ranker (OmniEmbed + RankVideo) improves performance, it still lacks the deep reasoning required for persona-constrained queries. Even the industry-leading Mixedbread model hits a performance ceiling at an nDCG@10 of 0.717, as it fundamentally relies on dense semantic overlap rather than cognitive logical alignment.

In contrast, our C2F-RAG pipeline achieves a striking nDCG@10 of 0.848, outperforming the strongest baseline by over 13 absolute percentage points. This exceptional ranking precision validates our core architectural hypothesis: pure semantic fetching is insufficient for massive collections. The breakthrough is directly attributable to the Adapted A.I.R. Agent. By performing deep logical reasoning over the SMC and applying a *hard-margin score calibration*, the system successfully penalizes semantic distractors and pushes true logical matches to the top 10 positions with near-zero position decay.

Method	nDCG@10	nDCG@100	R@10	R@100
OmniEmbed	0.166	0.245	0.096	0.297
OmniEmbed + RankVideo	0.542	0.546	0.423	0.494
Mixedbread	0.717	0.748	0.604	0.741
C2F-RAG (Ours)	**0.848**	**0.853**	**0.773**	**0.837**

Table 1: Retrieval performance on the MultiVENT 2.0 corpus. C2F-RAG significantly outperforms all baselines, demonstrating the necessity of cognitive logical filtering in large-scale multimodal retrieval.

	Information Generation (Info)			Temporal Grounding (Cite)			
Method	**P**	**R**	**F1**	**P**	**R**	**F1**	**Avg**
Baseline (CAG)	**0.651**	0.335	0.401	**0.510**	0.167	0.204	0.378
C2F-RAG (Ours)	0.557	**0.466**	**0.463**	0.452	**0.349**	**0.337**	**0.437**

Table 2: Quantitative results for Information Generation (Info) and Temporal Grounding (Cite) under the Oracle setting. The 'Avg' column represents the macro-average of the six preceding metrics. While the baseline achieves strong precision, our C2F-RAG system offers a more balanced precision-recall profile, leading to superior F1-scores and a higher overall average.

3.3 Generation and Grounding Performance

In this section, we evaluate the system's ability to synthesize persona-consistent responses and provide precise temporal grounding. We present the results under the ground-truth closed set (Oracle/Reference) evaluation setting, which isolates the generation performance from upstream retrieval variance. Table 2 summarizes the quantitative results against the official CAG baseline (Martin et al., 2025a), including the overall average metric (Avg) across all precision, recall, and F1-scores.

Analysis of Generation Performance Under the Oracle setting, C2F-RAG demonstrates highly competitive synthesis and grounding capabilities compared to the official CAG baseline. As shown in Table 2, our system achieves an Info F1-score of 0.463 and a Cite F1-score of 0.337, substantially outperforming the baseline metrics of 0.401 and 0.204, respectively. Furthermore, our system achieves an overall average score (Avg) of 0.437, demonstrating a robust generalization across all evaluation dimensions.

The Precision-Recall Balance A detailed examination of the metrics reveals distinct structural behaviors between the two systems. The CAG baseline adopts a high-precision approach, yielding strong precision scores in both Info (0.651) and Cite (0.510). However, this conservative generation strategy results in lower recall metrics.

C2F-RAG is designed to maximize information retention while strictly adhering to persona constraints. By utilizing our Prompt Sculpting mechanism, the system effectively extracts a broader spectrum of valid evidence from the video contexts. This architectural choice yields significant improvements in Information Recall (from 0.335 to 0.466) and Temporal Recall (from 0.167 to 0.349). Although this comprehensive extraction strategy incurs a slight reduction in absolute precision, the resulting balance significantly enhances both the task-specific F1-scores and the global average metric. This confirms that C2F-RAG successfully generates informative, well-grounded narratives without suffering from severe information omission.

3.4 Efficiency and Computational Cost

A critical requirement for Video RAG in production environments is computational feasibility. To evaluate the scalability of C2F-RAG, we report the inference latency using the `deepseek-chat` API with 15-thread parallel processing.

Fine-Grained Filtering Latency The Fine Stage represents the primary computational load as it involves evaluating 1,000 candidate videos per query. For the complete evaluation set (comprising 19 queries and totaling 19,000 video-query pairs), our system completed the cognitive reranking in approximately 4 minutes and 38 seconds using 15 parallel threads. This high throughput is primarily attributable to our *rapid-rejection logic*: for the vast majority of semantically similar but logically irrelevant distractors, the A.I.R. agent generates a null output ("None"). These negative cases typi-

Query	Persona	System Generated Output (Representative Snippets)	Citations
Q18	Cynical Journalist	"He reportedly lives in luxury, wearing a robe worth 160,000 RMB and driving an Audi luxury car... He has been dubbed the 'Shaolin CEO' for his commercial empire including Taobao shops and global trademarks."	`Xe2P8sYrT84` `2-FA5-LZtyI`
		"Sensational reports alleged he was arrested at Pudong Airport while attempting to flee to Los Angeles with 7 mistresses and 21 children, though these were later labeled as 'fake'."	`2-FA5-LZtyI`
Q19	Research Analyst	"On July 27, 2025, the Shaolin Temple's official website released a situation report alleging Shi Yongxin was involved in occupational embezzlement and misappropriation of funds."	`2-FA5-LZtyI` `BV1huSqB1EmS`
		"The Xinxiang People's Procuratorate approved the arrest of Shi Yongxin on November 16, 2025, for charges including non-state employee bribery and misappropriation of funds."	`BV1fmCQBWE2B` `8N4n_ArBp4A`

Table 3: Comparative synthesis of the same video evidence under different persona lenses from our final submission. The system successfully bifurcates the narrative logic based on user background.

cally terminate in under one second due to minimal token generation and simplified reasoning paths, allowing the system to prune massive candidate sets with remarkable efficiency.

Synthesis and Grounding Overhead In the final Generation Stage, the system synthesizes the distilled evidence into a persona-consistent response. The average processing time is 63 seconds per query. While higher than the filtering stage, this latency is consistent with the cognitive complexity required for multimodal long-context synthesis and the deterministic extraction of precise temporal timestamps. These results demonstrate that by decoupling the retrieval into a coarse-to-fine pipeline and leveraging multi-threaded parallelization, C2F-RAG achieves a scalable balance between deep reasoning and operational latency.

3.5 Case Study

To further illustrate the zero-shot persona adaptability of C2F-RAG, we present a qualitative comparison of the system's output for Query 18 and Query 19. Both queries involve the identical event—the 2025 Shi Yongxin controversy—but enforce conflicting persona constraints.

As shown in Table 3, the system adopts a sensational and texture-rich tone for the Journalist (Q18), highlighting luxury assets and controversial rumors. In contrast, for the Analyst (Q19), it suppresses these distractors and constructs a rigorous, institutional timeline grounded in legal terminology and official reports. This divergence confirms that the A.I.R. agent acts as a strict cognitive filter over the SMC, ensuring that the generated content is not just factually correct, but contextually relevant to the specific user bias (Zheng et al., 2023).

4 Conclusion

In this paper, we presented C2F-RAG, a fully training-free, cascaded Video RAG pipeline designed for large-scale multimodal augmented generation. By decoupling semantic pre-fetching from cognitive logical reasoning, our system effectively bridges the "semantic-logical gap" inherent in massive video collections. The architecture leverages BGE-M3 for high-recall coarse retrieval and an adapted A.I.R. agent for fine-grained cognitive reranking over the Serial Multimodal Context.

Experimental results on the MultiVENT 2.0 corpus demonstrate that C2F-RAG achieves a state-of-the-art nDCG@10 of 0.848 and an Info F1-score of 0.801, significantly outperforming existing baselines. Furthermore, our system exhibits remarkable robustness, maintaining superior generation quality even when the search space scales from a few dozen to over one hundred thousand videos. Future work will focus on improving fine-grained temporal grounding recall in open-set environments. C2F-RAG provides a scalable and economically viable plug-and-play solution for complex, persona-constrained multimodal reasoning tasks.

Acknowledgment

This work was supported by HUST Interdisciplinary Research Program under Grant No. 2025JCYJ077, the Ministry of Science and Technology of China under Grant No. 2025ZD0123800, and the KingSoft 2026 University-Industry Project.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *ICCV*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP*.

Reno Kriz, Kate Sanders, David Etter, Kenton Murray, Cameron Carpenter, Kelly Van Ochten, Hannah Recknor, Jimena Guallar-Blasco, Alexander Martin, Ronald Colaianni, Nolan King, Eugene Yang, and Benjamin Van Durme. 2025. Multivent 2.0: A massive multilingual benchmark for event-centric video retrieval. *Preprint*, arXiv:2410.11619.

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024. Open source strikes bread - new fluffy embeddings model.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*.

Yuning Lu, Jianzhuang Liu, Jian Yin, and Xinmei Tian. 2022. Learn to prompt for vision-language models. *IJCV*.

Xueguang Ma, Luyu Gao, Shengyao Zhuang, Jiaqi Samantha Zhan, Jamie Callan, and Jimmy Lin. 2025. Tevatron 2.0: Unified document retrieval toolkit across scale, language, and modality. *Preprint*, arXiv:2505.02466.

Sean MacAvaney, Andrew Craig, Craig Macdonald, and Iadh Ounis. 2022. ir-measures: Toward reproducible measures for information retrieval evaluation. In *European Conference on Information Retrieval*, pages 232–239. Springer.

Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836.

Alexander Martin, Reno Kriz, William Gantt Walden, Kate Sanders, Hannah Recknor, Eugene Yang, Francis Ferraro, and Benjamin Van Durme. 2025a. Wikivideo: Article generation from multiple videos. *Preprint*, arXiv:2504.00939.

Alexander Martin, William Walden, Reno Kriz, Dengjia Zhang, Kate Sanders, Eugene Yang, Chihsheng Jin, and Benjamin Van Durme. 2025b. Seeing through the mirage: Evaluating multimodal retrieval augmented generation. *Preprint*, arXiv:2510.24870.

Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau. 2021. Picard: Parsing incrementally for constrained auto-regressive decoding from language models. In *Proceedings of EMNLP*.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*.

Tyler Skow, Alexander Martin, Benjamin Van Durme, Rama Chellappa, and Reno Kriz. 2026. Rankvideo: Reasoning reranking for text-to-video retrieval. *Preprint*, arXiv:2602.02444.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *ICLR*.

Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Topic	Info F1		Cite F1	
	P	R	P	R
Average	55.7	46.6	45.2	34.9
Liberation_Day_Tariffs_q1	58.6	71.8	52.9	79.5
Shi_Yongxin_Scandal_q1	66.0	51.5	60.0	44.7
2025_Canadian_federal_election_q2	47.8	69.4	41.3	55.6
Blue_Ghost_Mission_1_q1	45.3	75.0	17.0	64.3
Liberation_Day_Tariffs_q2	50.0	59.0	42.0	59.0
Shi_Yongxin_Scandal_q2	56.0	51.5	66.0	51.5
2025_Alaskan_Typhoon_q1	71.4	42.9	7.1	0.0
2025_Myanmar_earthquake_q2	47.4	60.0	39.3	60.0
2025_Myanmar_earthquake_q1	40.0	73.3	40.0	40.0
2025_Alaskan_Typhoon_q2	68.1	41.3	4.3	0.0
Blue_Ghost_Mission_1_q2	38.8	64.3	36.7	50.0
Tropical_Storm_Wipha_q2	62.3	31.7	66.7	27.4
Palisades_Fire_q2	54.4	29.6	51.1	19.6
Tropical_Storm_Wipha_q1	77.8	24.6	75.0	20.3
Central_Texas_Floods_q1	59.0	26.7	56.4	15.6
Nepal_Youth_Protests_q1	59.6	23.5	60.9	14.7
2025_Canadian_federal_election_q1	23.5	52.8	24.7	38.9
Nepal_Youth_Protests_q2	72.4	20.6	65.5	14.7
Palisades_Fire_q1	60.3	16.9	52.4	7.9

Table 4: Detailed topic-level evaluation results of our proposed C2F-RAG system for Information Generation (Info F1) and Temporal Grounding (Cite F1) under the Oracle setting using the CLUE framework.

Dengjia Zhang, Alexander Martin, William Jurayj, Kenton Murray, Benjamin Van Durme, and Reno Kriz. 2026. Unified multimodal uncertain inference. *arXiv preprint arXiv:2604.08701*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*.

Yuanhao Zou, Shengji Jin, Andong Deng, Youpeng Zhao, Jun Wang, and Chen Chen. 2025. A.i.r.: Enabling adaptive, iterative, and reasoning-based frame selection for video question answering. *arXiv preprint arXiv:2510.04428*.

A Official MAGMaR Challenge Results (Topic-Level)

This section presents the comprehensive, topic-level evaluation results of our final system submission provided by the MAGMaR challenge organizers. While the main text reports the macro-averaged performance across all evaluation dimensions, Table 4 offers a fine-grained breakdown of Information Generation (Info) and Temporal Grounding (Cite) metrics for individual query topics under the Oracle/Reference setting evaluated via the CLUE framework.

B Detailed Implementation of Prompt Sculpting

The efficacy of C2F-RAG in navigating a 110,000-video corpus without prior fine-tuning relies heavily on the precision of our *Prompt Sculpting* mechanism. While the main body of this paper conceptualizes these processes as the "A.I.R. Agent" and the "Serial Multimodal Context (SMC)", this appendix provides the actual instructional logic used to operationalize these concepts. We decouple the reasoning process into two distinct stages: a logical pruning stage and a persona-consistent synthesis stage.

Prompt Template A: Cognitive Filtering Engine (Fine Stage)

```
SYSTEM INSTRUCTION:
You are an advanced, hyper-logical document retrieval and filtering engine. Your
    sole objective is to analyze a provided JSON database of video contexts (
    including visual descriptions, OCR text, and ASR transcripts) and score each
    video's logical alignment with a specific user query and persona.

--- PART 1: THE RELEVANCE SCORING MATRIX ---
You must score every single video candidate against the user's 'query' AND 'persona
    '. A video might be highly relevant to the general topic, but completely
    irrelevant to the persona (a "Hard Negative"). Assign a relevance score from
    0.00 to 1.00 using this strict rubric:

- [0.90 - 1.00] DIRECT & EXPLICIT ALIGNMENT: The video context contains explicit
    evidence that directly answers the core question of the query, perfectly
    matching the persona's needs.
- [0.75 - 0.89] STRONG CONTEXTUAL VALUE: Provides strong supporting information or
    secondary evidence related to the event.
- [0.50 - 0.74] TANGENTIAL / BACKGROUND RELEVANCE: Belongs to the correct event but
    provides little actionable intelligence for the specific persona.
- [0.30 - 0.49] WEAK / NOISY RELEVANCE: Barely related. Mentions the topic but
    focuses on different aspects.
- [0.00 - 0.29] IRRELEVANT (MUST BE EXCLUDED): Belongs to a different event or
    contains no useful information.

RULE: You must EXCLUDE any video with a score below 0.30. Sort the remaining videos
    in strictly descending order.

--- PART 2: STRICT OUTPUT SCHEMA ---
Output ONLY a raw JSON object. Do not use markdown formatting.
{
  "evaluations": [
    {
      "video_id": "<ID>",
      "reasoning": "<Justification based on the 5-tier rubric>",
      "relevance_score": 0.95
    }
  ]
}
```

Figure 2: Instructional skeleton for the Cognitive Filtering Agent. This stage focuses on logical pruning and score calibration.

B.1 Fine Stage: Cognitive Filtering and Hard Negative Pruning

As discussed in Section 2.3, the Fine Stage is designed to overcome the "semantic-logical gap" by simulating high-level human reasoning. The prompt for this stage (shown in Figure 2) is engineered to treat the SMC not merely as a text block, but as a multi-modal evidence database.

A key design element is the 5-tier Relevance Scoring Matrix. By forcing the LLM to categorize videos into granular bins (e.g., distinguishing "Strong Contextual Value" from "Tangential Background"), we enable the system to implement the Logic-Gated Exponential Attenuation (LGEA) formula with high confidence. This stage explicitly penalizes "Hard Negatives"—videos that pass initial semantic filters due to keyword overlap but fail the persona's logical requirements.

B.2 Generation Stage: Persona Adherence and Grounding

The final Generation Stage (Prompt B, Figure 3) shifts the focus from filtering to synthesis. The primary challenge here is maintaining the "Cognitive Lens" of the persona while ensuring zero-hallucination.

To achieve this, we implement two critical safeguards. First, the Zero Hallucination Directive explicitly forbids the injection of external world knowledge, a common failure mode in LLMs when dealing with famous entities or events. Second, the Persona Adaptation instructions force a lexicon shift, ensuring that a "Statistician" uses technical jargon while a "Journalist" focuses on narrative

Prompt Template B: Persona-Constrained Synthesis (Generation Stage)

```
SYSTEM INSTRUCTION:
You are an advanced, hyper-logical factual synthesis engine. You must act precisely
    as the persona described in the ‘background‘ field. Do not assume any external
    context or knowledge outside of the provided JSON database.

--- PART 1: LANGUAGE, BIAS, & PERSONA ADAPTATION ---
1. TARGET LANGUAGE OBLIGATION: You MUST write the generated text in the EXACT
    language specified by the user query (e.g., if "language": "nepali", you must
    use Nepali). JSON keys remain in English.
2. QUERY TYPE & BIAS CONTROL:
   - If "biased": Embrace the subjective agenda or emotional angle of the persona.
   - If "unbiased": Remain strictly objective, neutral, and analytical.
3. Lexicon and Jargon: Filter the context through the eyes of the persona. Use
    domain-specific terminology (e.g., a statistician uses "variance, vote share").

--- PART 2: THE ZERO HALLUCINATION DIRECTIVE ---
You are strictly forbidden from injecting world knowledge. If the user asks "How
    many seats did the party win?" and the database only says "The party won," you
    must state: "The video data confirms the win, but specific seat counts are
    unavailable." NEVER invent OCR text or statistics.

--- PART 3: GENERATION & CITATIONS ---
1. Sentence-level Citations: Every single sentence you write MUST be supported by
    the database. You must append a ‘citations‘ array to EVERY sentence object.
2. Global References: Provide a single, flat, deduplicated list of every ‘video_id‘
    cited.

--- PART 4: STRICT OUTPUT SCHEMA ---
Output exactly in the following JSON format. Do not use markdown blocks.
{
  "generation": {
    "responses": [
      {
        "text": "<Persona-aligned synthesized sentence in target language.>",
        "citations": ["<video_id>"]
      }
    ],
    "references": ["<video_id>"]
  }
}
```

Figure 3: Instructional skeleton for the Persona-Constrained Generator. This stage focuses on narrative synthesis, cross-lingual adaptability, and strict auditable grounding.

texture. Furthermore, the prompt enforces a strict JSON schema for *Deterministic Post-processing*, ensuring that the generated responses are perfectly formatted for automated evaluation and chunk-level temporal grounding.

MARQUIS: A Three-Stage Pipeline for Video Retrieval-Augmented Generation

Debashish Chakraborty[*2] **Dengjia Zhang**[*1] **Jialiang Jin**[*1]
Hanting Liu[*1] **Katherine Guerrerio**[*1] **Hanxiang Qin**[1] **Tyler Skow**[1]
Alexander Martin[1] **Reno Kriz**[1,2] **Benjamin Van Durme**[1,2]
[1]Johns Hopkins University
[2]Human Language Technology Center of Excellence
{dchakra6, amart233}@jhu.edu

Abstract

Retrieval-augmented generation from videos requires systems to retrieve relevant audiovisual evidence from large corpora and synthesize it into coherent, attributed text. Current approaches struggle at both ends: retrieval methods fail on complex, multi-faceted queries that cannot be captured by a single embedding, while generation methods lack the high-level reasoning needed to synthesize across multiple videos and face memory constraints over long, multi-video contexts. We present MARQUIS: a three-stage pipeline that addresses these limitations through (1) query expansion, fusion, and reranking, (2) calibrated structured evidence extraction, and (3) article generation from extracted evidence, optionally controlled by an RLM. On the MAGMaR2026 shared task, we improve retrieval performance from 0.195 to 0.759 (nDCG@10). For article generation, ITER-QA-BASE improves average human score from 3.09 to 3.83 over the CAG baseline, while MARQUIS-RLM achieves a human score of 3.30 and the strongest citation recall among non-QA systems.[1]

1 Introduction

Large-scale video corpora now document real-world events with a breadth and immediacy that no single text source can match, yet turning this raw audiovisual evidence into a well-sourced analytical article remains largely a manual process. Grounded article generation (Martin et al., 2025a) from large video collections requires systems to retrieve relevant audiovisual evidence and synthesize it into coherent, attributable text.

Current video retrieval and generation systems each face distinct limitations. Retrieval methods struggle with complex information needs that combine multiple implicit and explicit sub-needs and instructions in a single query (Weller et al., 2024), failing to surface all relevant videos to the information request. Generation methods face three interrelated challenges: long multi-video contexts exceed model memory constraints (Chen et al., 2024; He et al., 2024; Li et al., 2025), existing VLMs are not designed for multi-video reasoning, and most video understanding work remains focused on low-level recognition tasks like captioning and entity-centric QA rather than the high-level synthesis required for article generation (Martin et al., 2025a). These limitations compound in a full pipeline: retrieval errors propagate missing or irrelevant evidence into generation, where models already struggle to reason over the context they receive.

In this work, we present MARQUIS (**M**ultimodal **A**rticle generation via **R**etrieval, **Q**uery decomposition, **U**ncertainty calibration, and **I**terative evidence **S**ynthesis), a three-stage pipeline that addresses these limitations through modular decomposition of retrieval, evidence extraction, and generation. First, we decompose and expand each query into sub-queries, retrieve independently over each sub-query, and fuse the resulting ranked lists before reranking. Second, we extract evidence from retrieved videos through complementary query-agnostic and query-conditioned components, then calibrate each extracted claim against its source video to estimate support probability. Third, we generate cited articles from the curated evidence, comparing single-prompt, clustering-based, and bullet-point strategies. We additionally introduce MARQUIS-RLM, an instantiation of Recursive Language Models (RLM; Zhang et al., 2026a) that treats each pipeline module as a callable tool within a persistent structured-memory environment, enabling iterative evidence gathering, cross-video conflict resolution, and fact curation before generating the final article. Our contributions can be summarized as follows:

[*]Equal Contribution
[1]We release the code here: https://github.com/debashishc/marquis

Proceedings of the 2nd Workshop on Multimodal Augmented Generation via Multimodal Retrieval, pages 92–119
July 4, 2026

1. We introduce MARQUIS, a three-stage pipeline for large-scale video retrieval-augmented article generation.

2. Our two-stage retrieval approach, combining query expansion with rank fusion and video-native reranking, improves nDCG@10 from 0.195 to 0.759 over a dense retrieval baseline on MAGMaR2026.

3. Our QA-based article generation approach, combining query decomposition with video-grounded question answering, improves average human score from 3.09 to 3.83 over the CAG baseline on MAGMaR2026 oracle article generation.

2 Related Work

2.1 Multimodal Retrieval and RAG

Retrieval-augmented generation (RAG; Lewis et al., 2021) grounds language model outputs in retrieved evidence. Martin et al. (2025a) formalize multi-video article generation, the task of retrieval-augmented generation from multiple videos.

Retrieval Retrieving videos has been widely studied, but Kriz et al. (2025) show that most methods are specialized to descriptive queries and do not generalize to semantic queries or scale to large corpora. However, bi-encoder methods for dense (Luo et al., 2021; Zhu et al., 2024; Ma et al., 2025; Li et al., 2026), multi-vector (Reddy et al., 2025; Qin et al., 2026), and modality fusion (Samuel et al., 2025) provide scalable options for retrieval. Video reranking (Li et al., 2026; Skow et al., 2026) helps balance performance and scalability further, reranking the outputs of first-stage bi-encoder methods.

Generation Most work generating text from videos focuses on low-level extraction and single-video settings such as captioning and question answering (Xu et al., 2016; Krishna et al., 2017; Lei et al., 2018; Yu et al., 2019; Zhang et al., 2025). Martin et al. (2025a) show that existing VLMs fixate on low-level visual features and fail at the high-level synthesis required for article generation.

MARQUIS differs by integrating retrieval, calibrated extracted claims, QA-based evidence extraction, and iterative evidence control into a single system.

2.2 Long-Context Video Understanding

Recent long-video models extend temporal range through long-context training (Li et al., 2025), memory augmentation (He et al., 2024), and hierarchical compression (Chen et al., 2024), but still face computational and reasoning limits over extended multimodal context. Additionally, none of these methods are trained for multi-video settings. Recursive Language Models (RLMs; Zhang et al., 2026a) externalize long inputs into an interactive environment that can be inspected and processed through programmatic operations. We use this idea as a control layer for article generation: rather than placing all extracted evidence into one context, MARQUIS-RLM iteratively gathers, stores, and curates extracted claims before writing.

3 Retrieval

Our retrieval pipeline operates in two stages. First, we decompose each query into atomic sub-queries, retrieve independently over each, and fuse the resulting ranked lists into a single candidate set. Second, we rerank the fused candidates using a video-native reranker. Figure 1 illustrates the full pipeline.

3.1 First-Stage

Our first-stage method consists of three key components: (1) Query Decomposition and Expansion, (2) Dense Retrieval, and (3) Rank Fusion.

Query decomposition and expansion. We focus on queries that are long, instructional requests that combine a professional persona, domain background, and multi-faceted information need. Dense retrievers, however, are typically trained on short, single-intent queries, and encoding a complex request as a single vector collapses its many sub-needs into one point in embedding space. To bridge this gap, we decompose each query into N atomic sub-queries using an LLM, where each sub-query targets a single retrievable fact and is phrased as a concise search phrase.

Dense Retrieval. Both original queries and decomposed sub-queries are retrieved against an omnimodal index, which produces a ranked list of the top 1,000 candidates for each query.

Rank fusion. Given N ranked lists per query, we aggregate them into a single ranking. Let $\text{rank}(v, q_i)$ denote the rank of video v in the list produced by sub-query q_i, and let $s(v, q_i)$ denote

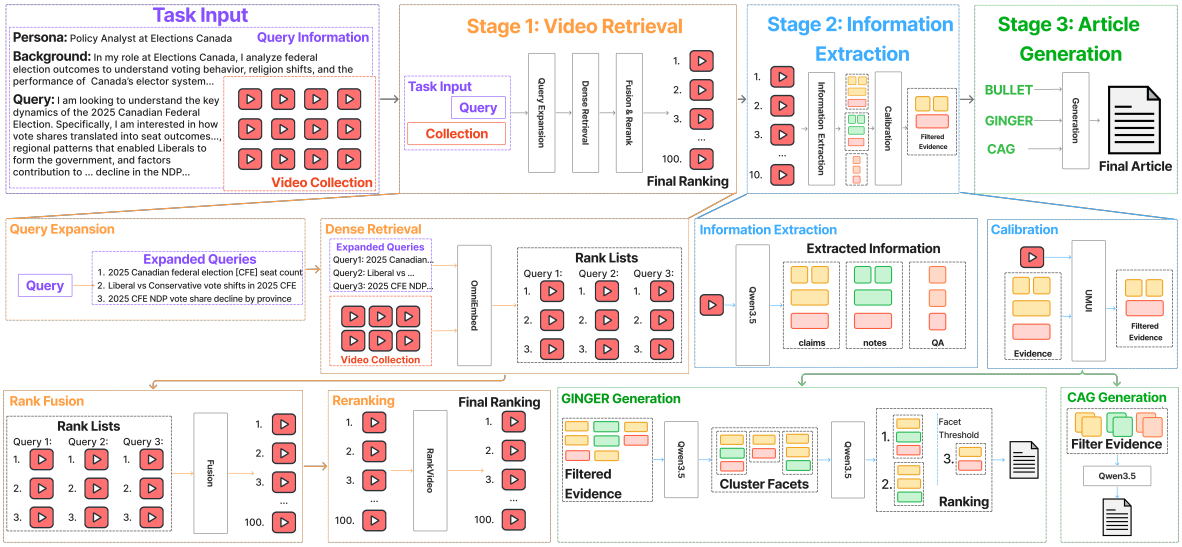


Figure 1: Overview of MARQUIS. **Stage 1 (Video Retrieval):** Each query is decomposed into sub-queries, which are independently encoded by OmniEmbed and retrieved against the corpus. The resulting ranked lists are fused and reranked by RANKVIDEO to produce the final ranking. **Stage 2 (Information Extraction):** Videos are processed by parallel information extraction streams—query-conditioned claims, query-agnostic notes, and QA—using Qwen3.5. Extracted evidence is scored by CLUE, a calibrated multimodal uncertain-inference model trained for the Unified Multimodal Uncertain Inference (UMUI) task, to filter unsupported claims. **Stage 3 (Article Generation):** The filtered evidence is then passed to an article generator: BULLET passes the list of extracted claims as the article, CAG single-prompt baseline, or GINGER-based generation to produce a final cited article.

the cosine similarity score. We evaluate five fusion strategies:

- **Reciprocal Rank Fusion (RRF).** Scores each video by the sum of reciprocal ranks across sub-query lists, with smoothing constant $K \in \{10, 60, 100\}$:

$$\text{RRF}_K(v) = \sum_{i=1}^{N} \frac{1}{K + \text{rank}(v, q_i)} \quad (1)$$

- **Sum of similarities.** Scores each video by the total cosine similarity across all sub-queries:

$$\text{Score}_{\text{sum}}(v) = \sum_{i=1}^{N} s(v, q_i) \quad (2)$$

- **Max similarity.** Scores each video by its highest similarity across sub-queries:

$$\text{Score}_{\text{max}}(v) = \max_i \; s(v, q_i) \quad (3)$$

- **Mean similarity.** Scores each video by the average similarity across sub-queries:

$$\text{Score}_{\text{mean}}(v) = \frac{1}{N} \sum_{i=1}^{N} s(v, q_i) \quad (4)$$

- **Weighted RRF.** Weights each reciprocal rank contribution by its cosine similarity:

$$\text{WRRF}_K(v) = \sum_{i=1}^{N} \frac{s(v, q_i)}{K + \text{rank}(v, q_i)} \quad (5)$$

Implementation details and expansion statistics are provided in Appendix C.

3.2 Reranking

Given the top 100 videos per query from the first-stage retrieval, we perform reranking with RankVideo (Skow et al., 2026). For each full query, we pass the candidate videos from the first-stage retrieval to RankVideo and reorder the ranked list based on these judgments.

4 Information Extraction

The retrieval stage returns video candidates, but article generation requires finer-grained evidence. We therefore convert retrieved videos into extracted claims that can be selected, calibrated, cited, and passed to the generation stage. Our evidence extraction system contains three components: *query-agnostic* note extraction, *query-conditioned* claim extraction, and question-answer extraction. The

components differ in how they condition on the query, but all produce source-linked extracted evidence with video identifiers and, when available, timestamps. The extraction component is shown at a high level in Figure 1; implementation details are provided in Figure 2.

Let $V(q)$ denote the videos associated with query q. For each video $v \in V(q)$, the extraction stage may produce three evidence families:

$$N(v) = \{n_1, \ldots, n_{|N(v)|}\},$$

$$C(v, q) = \{c_1, \ldots, c_{|C(v,q)|}\},$$

$$A(v, q) = \{a_1, \ldots, a_{|A(v,q)|}\},$$

where $N(v)$ denotes query-agnostic notes, $C(v, q)$ denotes query-conditioned claims, and $A(v, q)$ denotes question-answer outputs. Each output is later scored against its source video by the shared calibration stage.

4.1 Query-Agnostic Note Extraction

The query-agnostic component builds a reusable evidence base from each video without conditioning on a specific information need. Its goal is to capture directly observable visual events, on-screen text, and spoken content that may be useful across queries. The extractor is prompted to avoid speculation, causal inference, and cross-video synthesis. Each note describes a single atomic observation and includes a modality tag and optional timestamp. Confidence is not assigned at extraction time; support is estimated in a separate post-extraction calibration stage, which avoids conflating evidence extraction with support estimation.

4.2 Query-Conditioned Claim Extraction

The query-conditioned component targets evidence extraction toward a specific query. Where general notes prioritize breadth, this component prioritizes task relevance: given a specific information need, it extracts only claims that are relevant to that need. For each query-video pair, the extractor receives the query, persona, background, topic, and video metadata, and returns claims that are both query-relevant and directly supported by the video. This stage is not free-form answer generation: the prompt explicitly discourages generic scene descriptions, unsupported inferences, and redundant paraphrases. Each claim record contains a claim identifier, query identifier, video identifier, topic label, claim text, and optional support-oriented fields such as confidence, evidence description, source type, and timestamp. The resulting claims provide a targeted evidence set for downstream article generation.

4.3 Question-Answer Extraction

The question-answering component extracts evidence through targeted video question answering for information needs. Given a query, persona, and background, the system decomposes the information need into atomic subqueries, retrieves relevant videos for each subquery, and uses a vision-language model to answer using the retrieved video content and transcript. We implement two variants: a single-shot variant that answers a fixed set of decomposed subqueries and aggregates the grounded per-video answers without introducing external knowledge, and an iterative variant that generates follow-up questions conditioned on previous question-answer history to pursue missing or underspecified information. The output is a set of question-answer evidence records, each linked to the question, answer, source video, and any available timestamp or confidence metadata.

4.4 Video-Grounded Calibration

After extraction, each output is scored against its source video. Given a video v and artifact $x \in N(v) \cup C(v, q) \cup A(v, q)$, calibration produces a support probability

$$s_\theta(v, x) \in [0, 1].$$

The score estimates whether the output is supported by the source video. Calibration is applied after extraction so that evidence creation and support estimation remain separate. The calibrated outputs retain their original text and metadata, with the support score attached for downstream filtering and article generation. Implementation details and prompts are provided in Appendix E and Appendix I.

5 Article Generation

Our article generation pipeline synthesizes extracted video evidence into a fluent, source-attributed article that answers the information need. This stage operates after retrieval, evidence extraction, QA, and calibration. The article generator does not inspect raw videos directly, but instead consumes structured evidence artifacts tied to source videos and, when available, timestamps.

We design the article generation stage to be input-agnostic. In the experiments reported here,

the same generation procedures can operate over query-conditioned claims, query-agnostic notes, or question–answer pairs produced by the QA pipeline. Claims and notes include explicit video identifiers and timestamps, while QA pairs include the source videos used to produce the answer. The generator receives a flat list of extracted evidence together with their source metadata and is instructed to produce an article whose factual statements are supported by inline citations.

We compare three article generation strategies.

BULLET. The simplest strategy does not synthesize evidence into prose. It renders the retrieved evidence items directly as a numbered list of findings with inline citations. This output is conservative and preserves the connection between evidence and source videos, but it does not produce the coherent article-style response required by the task.

CAG is our Collaborative Article Generation baseline, adapted from WikiVideo (Martin et al., 2025a) to operate over extracted evidence. It generates a cited article from the extracted evidence for a query in one synthesis pass, following the role of the text-only aggregator in WikiVideo CAG. The model is instructed to organize evidence, remove redundancy, and preserve citations.

GINGER. We adapt the GINGER framework (Łajewska and Balog, 2025) to video-grounded extracted evidence. Since our extraction stage already produces atomic evidence, we skip nugget detection and perform facet clustering, cluster ranking, per-cluster summarization, and fluency enhancement. The model first groups evidence into thematic clusters (e.g., casualties, rescue efforts, government response), ranks them by query relevance, summarizes the top clusters independently into short cited sentences, and finally rewrites them into a coherent article. This staged-decision decomposes article generation into smaller controlled calls and helps preserve citations.

6 RLM Controller

In addition to the aforementioned pipeline, we further construct a RLM-based high-level control system for article generation, MARQUIS-RLM, organizing each module of the pipeline as a tool to be called by the Root LM.

Conceptually, RLM serves here as a general recursive control paradigm, whereas MARQUIS-RLM is our task-specific instantiation of this paradigm for multi-video article generation. Unlike standard code-generating RLM, MARQUIS-RLM equips Root LM to call our pre-developed modules in REPL, preserving robust performance of specialized modules while gaining the reasoning and efficiency of the RLM paradigm. We further make explicit structured memory a core component of the system: Root LM always reasons over an evidence record that can be searched, reused, and revised, rather than relying only on what remains in context. This mitigates the information forgetting and multi-source confusion common in long iterative workflows, while making cross-video conflicts and missing information explicit. Examples are provided in Appendix H.

REPL Environment and Tool Interface. We instantiate MARQUIS-RLM in a persistent Python sandbox whose namespace contains task context, memory bank, and callable sub-tools adapted from the modules in previous sections. At each iteration, the Root LM generates and executes code in the persistent namespace, accesses raw video, audio, and transcripts only through callable tools.

Memory Bank. Under MAGMaR's long-context setting, the Root LM could face recurring failures including evidence forgetting, cross-source confusion, and missed conflicts or information gaps. All stem from the same limitation: the Root LM can only reason over what remains visible in context (Shi et al., 2026; Zhang et al., 2026c). Even the original free-form RLM-style REPL state is insufficient, since it introduces naming drift, reassignment errors, schema drift, and perception-level confusion. To address this, inspired by recent external-memory designs for LM agents such as MemR^3 (Du et al., 2025), we build a dynamic structured memory on top of the REPL.

The full schema and operators are given in Appendix H.

Think–Act–Observe. We require the Root LM to follow a coarse-grained Think–Act–Observe mechanism, inspired by interleaved reasoning-and-acting frameworks such as ReAct (Yao et al., 2023).

This design enforces immediate external feedback at each step and grounds reasoning in explicit state transitions, while still leaving tool choice and reflection frequency to the Root LM.

	Run	nDCG@10	nDCG@20	nDCG@100	R@10	R@20	R@100
	OmniEmbed	0.195	0.229	0.311	0.190	0.276	0.494
Sim	Max Sim	**0.722**	**0.743**	**0.784**	**0.639**	0.731	0.826
	Mean Sim	0.637	0.650	0.696	0.544	0.618	0.736
	Sum Sim	0.703	0.725	0.776	0.604	0.698	0.818
RRF	K=10	0.700	0.739	0.777	0.612	**0.735**	**0.832**
	K=60	0.695	0.728	0.773	0.599	0.714	0.823
	K=100	0.688	0.719	0.767	0.590	0.704	0.818
	Weighted	0.699	0.730	0.778	0.604	0.714	**0.832**

Table 1: First-stage retrieval results. Best score per column is **bolded**.

7 Experiments

Dataset We evaluate on the MAGMaR2026 Test Set. The data is based on WikiVideo (Martin et al., 2025a). For the retrieval and RAG settings, we retrieve relevant videos from a combination of MAGMaR and MultiVENT2.0 (Kriz et al., 2025). For oracle article generation, systems receive the ground-truth relevant videos, isolating generation quality from retrieval quality.

Evaluation Setup Retrieval results are evaluated with nDCG and Recall for 10, 20, and 100. We use the ir-measures (MacAvaney et al., 2022) to calculate these scores. Generated articles are evaluated with an automatic and human evaluation. For the automatic evaluation, the systems are evaluated by MiRAGE (Martin et al., 2025b), which captures the factuality, information coverage, groundedness, and proper attribution of citations. Each MiRAGE entailment judgment is judged by CLUE (Zhang et al., 2026b). For human evaluation, three human annotators provide scalar scores of 1–5 for each system, scoring factuality, adequacy, coherence, relevancy, and fluency. After providing scalar scores for each prediction, the annotators also pick the best system response to each query.

Experimental Setup. All systems are evaluated under the same MAGMaR2026 retrieval and oracle-generation splits, but differ in their video access patterns and model backends. Retrieval uses OmniEmbed (Ma et al., 2025) for video and query encoding and Qwen3.5-9B (Team, 2026) for query decomposition. The information-extraction streams (note and claim extraction, calibration) use Qwen3.5-9B over sampled video frames. The QA pipeline combines Qwen3.5-27B for answer generation, Qwen2.5-Omni-7B (Xu et al., 2025) with OmniEmbed for multimodal embeddings, and Whisper medium.en (Radford et al., 2023) for transcription. Article generation uses Qwen3.5-27B for both the single-prompt baseline and GINGER-based generators. The RLM controller runs a Qwen3.5-9B root LM that calls the extraction and QA modules as sub-tools. None of the claim-based extraction or generation systems use audio; only the QA pipeline and RLM (via its transcription tool) access the audio stream. Frame rates, top-k values, generation budgets, and other component-specific hyperparameters are listed in Appendix A.

7.1 Retrieval

In Table 1 and Table 2, we report the results of video retrieval for first-stage and reranking, respectively. All query expansion and fusion methods substantially outperform the OmniEmbed dense retrieval baseline. This confirms that decomposing complex queries into sub-queries targeting atomic pieces of information is much more suitable for a dense retriever. This is an intuitive result, as most first-stage retrievers are trained on short, single-intent query-document pairs and compressing a complex information request into a single embedding is out-of-distribution and challenging (Weller et al., 2024). However, our sub-queries reduce this burden, allowing for the model to interface with in-distribution queries and leaving the merging of those ranked lists to a fusion or reranking approach.

Similarity vs. RRF fusion. Among first-stage methods (Table 1), Max similarity achieves the highest nDCG at all cutoffs. It benefits from its selection mechanism: because it scores each video by its best-matching sub-query, it surfaces videos that are highly relevant to at least one facet of the information need, even if they are irrelevant to oth-

	Run	nDCG@10	nDCG@20	nDCG@100	R@10	R@20	R@100
	OmniEmbed + RV	0.542	0.534	0.546	0.423	0.462	0.494
		177.95	133.19	75.56	122.63	67.39	N/A
Sim	Max Sim + RV	0.399	0.405	0.425	0.344	0.383	0.437
		-44.74	-45.49	-45.79	-46.17	-47.61	-47.09
	Mean Sim + RV	0.740	0.723	0.750	0.637	0.665	0.736
		16.17	11.23	7.76	17.10	7.61	N/A
	Sum Sim + RV	0.747	0.758	0.800	0.636	0.711	0.818
		6.26	4.55	3.09	5.30	1.86	N/A
RRF	RRF K=10 + RV	**0.759**	**0.771**	**0.811**	**0.652**	**0.735**	**0.832**
		8.43	4.33	4.38	6.54	N/A	N/A
	RRF K=60 + RV	0.754	0.765	0.807	0.641	0.716	0.823
		8.49	5.08	4.40	7.01	0.28	N/A
	RRF K=100 + RV	0.746	0.757	0.799	0.636	0.711	0.818
		8.43	5.29	4.17	7.80	0.99	N/A
	Weighted RRF + RV	0.757	0.768	0.810	0.650	0.725	**0.832**
		8.30	5.21	4.11	7.62	1.54	N/A

Table 2: Reranked retrieval results with percentage change relative to first-stage baseline. Green denotes improvement, red denotes degradation.

ers. RRF strategies (Cormack et al., 2009), which aggregate evidence across all sub-queries, produce more balanced rankings and achieve higher recall at deeper cutoffs, suggesting they are better at capturing the full breadth of a multi-faceted query. Mean and Sum similarity consistently underperforms the other aggregation methods, likely because averaging dilutes strong matches with weak ones. Among the RRF variants, lower K values perform slightly better, as a smaller smoothing constant amplifies rank differences and rewards videos that appear near the top of multiple sub-query lists. Weighted RRF performs comparably to standard RRF, indicating that weighting reciprocal ranks by cosine similarity provides limited additional signal when the sub-queries are already well-targeted.

Reranking. As shown in Table 2, applying RankVideo reranking improves performance across nearly all fusion strategies. Among the expanded-query methods, all RRF variants and similarity variants (except Max) see consistent improvements, with RRF at K=10 achieving the best ranking performance overall. The one notable exception is Max similarity, where reranking sharply degrades all metrics. We leave a detailed analysis of this failure mode to future work.

7.2 Generation

In Table 3, we report oracle generation results, where each system receives the ground-truth relevant videos rather than retrieved candidates, isolating generation quality from retrieval effects. We evaluate eight system variants spanning three evidence pipelines: claim-based extraction (BULLET, Ginger), question answering (iterative (Iter QA) and single shot (SS QA)), and RLM-controlled generation.

Generation Systems. Among the claim-based generation variants, GINGER is the strongest prose generator, improving over the CAG baseline in human score, best-vote share, and both information and citation precision. Its staged decomposition into facet clustering, ranking, and per-cluster summarization appears to help the model organize evidence and preserve citations more reliably than a single generation call. BULLET shows the opposite tradeoff: it achieves slightly higher citation recall than the other claim-based systems, but receives the lowest human score and no best-system votes, confirming that annotators penalize outputs that lack fluent synthesis even when source attribution is preserved. Taken together, these results suggest that explicit topical organization improves generation quality, but that the final output must still read as coherent prose to satisfy analyst information needs.

System	Human Score	Best Votes	Best %	Info		Cite	
				P	R	P	R
CAG (baseline)	3.09	1	1.8%	76.4	**41.0**	61.7	22.8
BULLET	2.67	0	0.0%	71.1	39.4	60.4	23.7
GINGER	3.12	6	10.5%	**77.6**	40.4	**64.3**	22.6
MARQUIS-RLM	3.30	3	5.3%	70.8	38.5	59.2	27.2
SS QA Base	3.07	6	10.5%	33.1	30.6	27.7	**28.1**
SS-QA-GINGER	3.42	**10**	**17.5%**	54.4	32.4	32.6	23.8
ITER-QA-BASE	**3.83**	8	14.0%	34.7	31.3	26.8	25.8
ITER-QA-GINGER	3.69	5	8.8%	34.5	29.0	25.7	22.6

Table 3: Oracle generation results for MARQUIS systems. H = human score; B = best-system votes; IP/IR = information precision/recall; CP/CR = citation precision/recall.

QA Systems. QA-based systems achieve the strongest human preference scores. ITER-QA-BASE obtains the highest average human score, while SS-QA-GINGER receives the most best-system votes. Their aggregate automatic metrics are weaker, largely because QA failures on a small number of topics produce conservative empty or near-empty outputs. This suggests that QA improves article usefulness when relevant answers are recovered, but remains brittle when decomposition or video-level answering fails. When the QA systems fail to answer questions, due to VLM failures or irrelevant sub-questions, the downstream generation systems often refuse to write the article, backing off due to insufficient evidence. This conservative behavior is a double-edged sword: it avoids hallucination on topics where the video evidence genuinely lacks the requested information (e.g., Myanmar Earthquake Q1), but it produces zero-score outputs that sharply deflate aggregate metrics on topics where information was available.

RLM. MARQUIS-RLM improves human score over CAG and BULLET and achieves the highest citation recall among non-QA systems. This suggests that iterative evidence gathering and structured memory help preserve attribution across multi-video contexts. Its lower precision and citation precision, however, indicates that the controller also admits less relevant facts into the final article. We therefore view MARQUIS-RLM as an evidence-management mechanism rather than a standalone replacement for structured generation; its Think–Act–Observe loop and persistent memory bank are effective at resolving cross-video conflicts and filling information gaps (see examples in Appendix H), and tighter integration with GINGER-based synthesis is a natural next step.

8 Conclusion

We presented MARQUIS, a three-stage pipeline for video retrieval-augmented article generation. MARQUIS decomposes complex queries, retrieves and reranks relevant videos, converts video content into calibrated extracted evidence, and generates cited articles from selected evidence. The optional MARQUIS-RLM controller extends this pipeline by treating retrieval, extraction, QA, calibration, and generation as tools within a structured-memory environment, enabling iterative evidence gathering and curation before writing. Our experiments show that explicit query decomposition and video-native reranking substantially improve retrieval, while article-generation results reveal complementary tradeoffs among claim-based, QA-based, and RLM-controlled systems. More broadly, our findings suggest that grounded generation from video is best framed as an evidence-management problem. Rather than prompting a model to summarize long multi-video context directly, effective systems should retrieve broadly, extract atomically, estimate source support, and synthesize only from selected extracted evidence. Future work should improve learned calibration, integrate retrieval and extraction more tightly, and develop generation methods that combine structured evidence organization with iterative evidence control.

Acknowledgment

This material is based upon work supported by the NSF Graduate Research Fellowship under Grant No. DGE2139757. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. 2024. Longvila: Scaling long-context visual language models for long videos. *Preprint*, arXiv:2408.10188.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Xingbo Du, Loka Li, Duzhen Zhang, and Le Song. 2025. MemR^3: Memory retrieval via reflective reasoning for llm agents. *Preprint*, arXiv:2512.20237.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. MA-LMM: Memory-augmented large multimodal model for long-term video understanding. *Preprint*, arXiv:2404.05726.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. *Preprint*, arXiv:1705.00754.

Reno Kriz, Kate Sanders, David Etter, Kenton Murray, Cameron Carpenter, Kelly Van Ochten, Hannah Recknor, Jimena Guallar-Blasco, Alexander Martin, Ronald Colaianni, Nolan King, Eugene Yang, and Benjamin Van Durme. 2025. MultiVENT 2.0: A massive multilingual benchmark for event-centric video retrieval. *Preprint*, arXiv:2410.11619.

Weronika Łajewska and Krisztian Balog. 2025. Ginger: Grounded information nugget-based generation of responses. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 2723–2727, New York, NY, USA. Association for Computing Machinery.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A Unified Framework for State-of-the-Art Multimodal Retrieval and Ranking. *Preprint*, arXiv:2601.04720.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. 2025. VideoChat-Flash: Hierarchical compression for long-context video modeling. *Preprint*, arXiv:2501.00574.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2021. CLIP4Clip: An empirical study of clip for end to end video clip retrieval. *Preprint*, arXiv:2104.08860.

Xueguang Ma, Luyu Gao, Shengyao Zhuang, Jiaqi Samantha Zhan, Jamie Callan, and Jimmy Lin. 2025. Tevatron 2.0: Unified document retrieval toolkit across scale, language, and modality. *Preprint*, arXiv:2505.02466.

Sean MacAvaney, Craig Macdonald, and Iadh Ounis. 2022. Streamlining evaluation with ir-measures. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II*, volume 13186 of *Lecture Notes in Computer Science*, pages 305–310. Springer.

Alexander Martin, Reno Kriz, William Gantt Walden, Kate Sanders, Hannah Recknor, Eugene Yang, Francis Ferraro, and Benjamin Van Durme. 2025a. WikiVideo: Article generation from multiple videos. *Preprint*, arXiv:2504.00939.

Alexander Martin, William Walden, Reno Kriz, Dengjia Zhang, Kate Sanders, Eugene Yang, Chihsheng Jin, and Benjamin Van Durme. 2025b. Seeing Through the MiRAGE: Evaluating Multimodal Retrieval Augmented Generation. *Preprint*, arXiv:2510.24870.

Hanxiang Qin, Alexander Martin, Rohan Jha, Chunsheng Zuo, Reno Kriz, and Benjamin Van Durme. 2026. Multi-Vector Index Compression in Any Modality. *Preprint*, arXiv:2602.21202.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of

Proceedings of Machine Learning Research, pages 28492–28518. PMLR.

Arun Reddy, Alexander Martin, Eugene Yang, Andrew Yates, Kate Sanders, Kenton Murray, Reno Kriz, Celso M. de Melo, Benjamin Van Durme, and Rama Chellappa. 2025. Video-ColBERT: Contextualized late Interaction for Text-to-Video Retrieval. *Preprint*, arXiv:2503.19009.

Saron Samuel, Dan DeGenaro, Jimena Guallar-Blasco, Kate Sanders, Oluwaseun Eisape, Tanner Spendlove, Arun Reddy, Alexander Martin, Andrew Yates, Eugene Yang, Cameron Carpenter, David Etter, Efsun Kayi, Matthew Wiesner, Kenton Murray, and Reno Kriz. 2025. MMMORRF: Multimodal Multilingual Modularized Reciprocal Rank Fusion. *Preprint*, arXiv:2503.20698.

Yaorui Shi, Yuxin Chen, Siyuan Wang, Sihang Li, Hengxing Cai, Qi Gu, Xiang Wang, and An Zhang. 2026. Look back to reason forward: Revisitable memory for long-context llm agents. *Preprint*, arXiv:2509.23040.

Tyler Skow, Alexander Martin, Benjamin Van Durme, Rama Chellappa, and Reno Kriz. 2026. Rankvideo: Reasoning reranking for text-to-video retrieval. *Preprint*, arXiv:2602.02444.

Qwen Team. 2026. Qwen3.5-omni technical report. *Preprint*, arXiv:2604.15804.

Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. 2024. FollowIR: Evaluating and teaching information retrieval models to follow instructions. *Preprint*, arXiv:2403.15246.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. ActivityNet-QA: A dataset for understanding complex web videos via question answering. *Preprint*, arXiv:1906.02467.

Alex L. Zhang, Tim Kraska, and Omar Khattab. 2026a. Recursive Language Models. *Preprint*, arXiv:2512.24601.

Dengjia Zhang, Alexander Martin, William Jurayj, Kenton Murray, Benjamin Van Durme, and Reno Kriz. 2026b. Unified multimodal uncertain inference. *Preprint*, arXiv:2604.08701.

Dengjia Zhang, Charles Weng, Katherine Guerrerio, Yi Lu, Kenton Murray, Alexander Martin, Reno Kriz, and Benjamin Van Durme. 2025. HLTCOE Evaluation Team at TREC 2025: VQA Track. *Preprint*, arXiv:2512.07738.

Yuxiang Zhang, Jiangming Shu, Ye Ma, Xueyuan Lin, Shangxi Wu, and Jitao Sang. 2026c. Memory as Action: Autonomous context curation for long-horizon agentic tasks. *Preprint*, arXiv:2510.12635.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. 2024. LanguageBind: Extending video-language pretraining to N-modality by language-based semantic alignment. *Preprint*, arXiv:2310.01852.

A Experimental Setup Details

This appendix summarizes the model backends, input access, and hyperparameters used by each component. We report model backends and hyperparameters for retrieval, information extraction, calibration, and QA in Table 4, and article generation hyperparameters in Table 5. Prompt templates are listed in Appendix I.

B Additional Results

Tables 6, 7, 8, 9, and 10 report per-query scores across all systems and topics. Claim-based systems (CAG, Bullet, Ginger) consistently achieve high information precision but lower recall. QA systems suffer catastrophic zero-score failures on several topics, including the Alaskan Typhoon (Table 6a), Central Texas Floods (Table 8a), Nepal Youth Protests (Table 9a), and Shi Yongxin Scandal (Table 10a), where the QA pipeline fails to retrieve relevant videos and the generator backs off rather than hallucinate. The RLM performs most distinctively on the Canadian Federal Election (Table 6b), where cross-video conflict resolution yields the highest citation precision and recall on Q2, and on Nepal Youth Protests, where it substantially outperforms all other systems. The Palisades Fire (Table 9b) and Tropical Storm Wipha (Table 10b) exhibit uniformly high precision but very low recall across all systems, suggesting broad reference sets that no system fully covers.

In Table 11 and Table 12 we report the overall rankings of each system from the MAGMaR shared

Component	Backend	Setting	Value
Retrieval	OmniEmbed; RankVideo	Decomposition model	Qwen3.5-27B (T=0.7, top-p=0.9, 2048 tok)
		Thinking mode	Disabled
		Embedding pooling / norm	End-of-sequence; L2
		Precision	bfloat16
		Corpus size	109,814 videos
		First-stage depth	100 videos per (sub-)query
		Fusion methods	Max, Mean, Sum, RRF, WRRF ($K \in \{10, 60, 100\}$)
		Reranking depth	100 videos
Note / Claim Extraction	Qwen3.5-9B	FPS / max frames	1.0 / 128
		Decoding	T=0.3, top-p=0.8, top-k=20
		Max tokens (notes / claims)	2048 / 4096
		Seed (notes / claims)	42 / 40
		Thinking	off
Calibration	CLUE; prompted Qwen3.5	FPS	0.5
		Frame size	256×256
		Filtering threshold	0.5
QA	Qwen3.5-27B; OmniEmbed; Whisper medium.en	Questions / query	10–25
		Question decoding	T=0.4, top-p=0.9, 1024 tok
		Video QA / aggregation tokens	512 / 256
		Iterative max steps	5 / question
		Frame budget	32 frames
		Audio	16 kHz mono
RLM	Qwen3.5-9B (root, VLM); Qwen3.5-27B (judge); tools as above	Root LM context	32,768 tokens
		Caption VLM	32 frames, 32,000 tok
		Caption VLM decoding	T=0.3
		Max iterations	60
		LLM-as-a-Judge	T=0.2, 512 tok, per-iteration

Table 4: Unified component backends and hyperparameters for all MARQUIS pipeline stages.

Method	Input	Max Tokens	Decoding
BULLET	Selected claims	–	–
GINGER cluster	Claims	2048	T=.3, p=.9
GINGER rank	Clusters	512	T=.3, p=.9
GINGER summarize	Top 5 clusters	256	T=.5, p=.9
GINGER fluency	Summaries	1024	T=.7, p=.9

Table 5: Article generation hyperparameters. All methods use Qwen3.5-27B.

task leaderboards for retrieval and generation, respectively. Our retrieval systems hold the 2nd-6th place positions. Our generation systems place 1st and 3rd-6th.

C Appendix: Retrieval Implementation and Full Ablation

This appendix consolidates the retrieval implementation details, query expansion artifacts, retrieval figure, ablations, and hyperparameters. Prompt templates for query decomposition are listed in Appendix I.

C.1 Query and Corpus Encoding

Each MAGMaR query is represented by concatenating the persona title, background, and query text. The original queries and generated sub-queries are encoded with OmniEmbed using the same query prefix, appended end-of-text token, end-of-sequence pooling, and L2 normalization. The video corpus contains 109,814 videos, consisting of 109,724 MultiVENT 2.0 videos and 90 MAGMaR2026 test videos. Search uses cosine similarity over normalized embeddings and returns the top 100 videos per query or sub-query.

C.2 Sub-query Expansion Statistics

The final flattened sub-query file contains 430 sub-queries across 19 original queries, for an average of 22.63 sub-queries per query. The minimum is 1 and the maximum is 25. The minimum is caused by one malformed decomposition output that fell back to a single query-like search probe in the flattened retrieval file. Excluding this fallback case, the 18

	System	Info F1		Cite F1	
		P	R	P	R
Q1	CAG	**94.1**	**33.3**	2.9	0.0
	Bullet	77.3	28.6	**9.1**	0.0
	Ginger	91.4	**33.3**	8.6	0.0
	RLM	81.5	28.6	3.7	0.0
	Iter-B	0.0	0.0	0.0	0.0
	Iter-G	0.0	0.0	0.0	0.0
	SS-B	19.4	3.2	0.0	0.0
	SS-G	27.3	4.8	2.3	0.0
Q2	CAG	87.5	**41.3**	4.2	0.0
	Bullet	**94.7**	34.9	0.0	0.0
	Ginger	87.5	**41.3**	0.0	0.0
	RLM	71.4	38.1	3.6	0.0
	Iter-B	0.0	0.0	0.0	0.0
	Iter-G	0.0	0.0	0.0	0.0
	SS-B	8.8	6.3	6.1	0.0
	SS-G	35.3	3.2	**12.0**	0.0

(a) 2025 Alaskan Typhoon

	System	Info F1		Cite F1	
		P	R	P	R
Q1	CAG	**57.1**	36.1	33.3	13.9
	Bullet	18.5	41.7	18.5	27.8
	Ginger	50.0	36.1	28.1	13.9
	RLM	27.3	44.4	21.2	**50.0**
	Iter-B	37.0	**52.8**	**44.4**	47.2
	Iter-G	22.2	44.4	27.8	47.2
	SS-B	31.2	38.9	28.6	38.9
	SS-G	38.5	33.3	38.5	38.9
Q2	CAG	43.4	61.1	42.3	33.3
	Bullet	46.7	41.7	46.7	13.9
	Ginger	43.6	61.1	43.6	33.3
	RLM	**64.7**	58.3	**70.6**	52.8
	Iter-B	43.4	61.1	43.4	58.3
	Iter-G	39.6	44.4	38.5	38.9
	SS-B	55.0	69.4	48.3	**69.4**
	SS-G	58.7	**75.0**	58.7	63.9

(b) 2025 Canadian Federal Election

Table 6: Per-query scores across all systems, judge: CLUE. Iter-B/G = Iter-QA-Base/Ginger, SS-B/G = SS-QA-Base/Ginger.

successfully decomposed queries produce 429 sub-queries, with a minimum of 22, an average of 23.83, and a maximum of 25 sub-queries per query.

C.3 Qualitative Expansion Examples

In Table 14 we show some example expansions from our method.

C.4 Fusion and Reranking

We evaluate both score-based and rank-based fusion over the sub-query ranked lists. The score-based methods are sum similarity, max similarity, and mean similarity. The rank-based methods are reciprocal rank fusion with $K \in \{10, 60, 100\}$ and weighted reciprocal rank fusion, where reciprocal-rank contributions are weighted by cosine similarity. We also evaluate reranked variants in which the top 100 first-stage candidates are reordered with RANKVIDEO.

C.5 Dropping Sub-queries

To test whether query decomposition depends on a small number of strong sub-queries or on broad facet coverage, we randomly retain only k sub-queries per original query before fusion, for $k \in \{1, 5, 10\}$. We repeat each random condition over five seeds and report mean and standard deviation. The full system uses all generated sub-queries.

Performance improves monotonically as more sub-queries are retained. A single random sub-query already outperforms the no-expansion baseline, showing that the decomposition often produces useful search probes. However, retaining 5 or 10 sub-queries substantially improves both ranking quality and recall, and using all sub-queries gives the best overall performance. This suggests that the gains from decomposition come not only from finding one strong reformulation but also from covering multiple facets of the original query.

D Appendix: Information Extraction Implementation

This appendix provides implementation details for the information extraction stage, including general note extraction, query-conditioned claim extraction, artifact schemas, representative outputs, and query–topic alignment. Prompt templates for note extraction, claim extraction, and calibration are listed in Appendix I. Figure 2 illustrates the information extraction and calibration workflow.

	System	Info F1		Cite F1	
		P	R	P	R
	CAG	60.5	**93.3**	55.8	80.0
Q1	Bullet	69.8	**93.3**	40.0	0.0
	Ginger	**82.9**	86.7	**74.3**	**86.7**
	RLM	68.4	60.0	54.1	46.7
	Iter-B	45.2	60.0	40.5	33.3
	Iter-G	38.3	60.0	27.7	46.7
	SS-B	17.6	40.0	15.7	40.0
	SS-G	30.2	53.3	14.0	60.0
	CAG	**78.1**	**93.3**	63.3	80.0
Q2	Bullet	61.9	86.7	**71.4**	80.0
	Ginger	76.5	**93.3**	60.6	80.0
	RLM	73.3	80.0	63.3	**86.7**
	Iter-B	68.4	86.7	59.6	66.7
	Iter-G	54.3	80.0	48.9	73.3
	SS-B	61.8	80.0	44.4	80.0
	SS-G	68.6	86.7	61.2	73.3

(a) 2025 Myanmar Earthquake

	System	Info F1		Cite F1	
		P	R	P	R
	CAG	88.6	42.9	71.4	21.4
Q1	Bullet	**95.5**	39.3	77.3	35.7
	Ginger	88.2	42.9	**79.4**	21.4
	RLM	63.0	42.9	55.6	42.9
	Iter-B	48.4	64.3	43.5	**67.9**
	Iter-G	63.5	**71.4**	57.7	46.4
	SS-B	62.9	67.9	44.9	64.3
	SS-G	72.1	67.9	63.4	**67.9**
	CAG	77.4	57.1	67.7	17.9
Q2	Bullet	72.2	57.1	55.6	46.4
	Ginger	**80.8**	57.1	**69.2**	17.9
	RLM	66.7	35.7	66.7	35.7
	Iter-B	30.3	71.4	22.7	**71.4**
	Iter-G	42.2	64.3	31.1	53.6
	SS-B	29.9	**75.0**	25.4	67.9
	SS-G	47.5	57.1	35.0	28.6

(b) Blue Ghost Mission 1

Table 7: Per-query scores across all systems, judge: CLUE. Iter-B/G = Iter-QA-Base/Ginger, SS-B/G = SS-QA-Base/Ginger.

	System	Info F1		Cite F1	
		P	R	P	R
	CAG	65.5	**20.0**	61.1	11.1
Q1	Bullet	56.2	15.6	56.2	**13.3**
	Ginger	66.7	**20.0**	**71.4**	6.7
	RLM	43.2	13.3	35.1	**13.3**
	Iter-B	0.0	0.0	0.0	0.0
	Iter-G	0.0	0.0	0.0	0.0
	SS-B	1.7	2.2	1.7	4.4
	SS-G	**72.1**	15.6	1.5	0.0

(a) Central Texas Floods

	System	Info F1		Cite F1	
		P	R	P	R
	CAG	70.0	69.2	63.2	56.4
Q1	Bullet	63.6	61.5	59.1	59.0
	Ginger	63.2	69.2	54.1	56.4
	RLM	44.8	61.5	41.4	38.5
	Iter-B	60.9	**76.9**	47.8	71.8
	Iter-G	75.0	**76.9**	66.7	69.2
	SS-B	64.6	74.4	64.6	**76.9**
	SS-G	**75.9**	64.1	**67.9**	59.0
	CAG	72.4	48.7	69.0	43.6
Q2	Bullet	64.3	51.3	57.1	48.7
	Ginger	67.7	48.7	**71.0**	43.6
	RLM	**75.0**	53.8	62.5	41.0
	Iter-B	51.6	53.8	39.1	**56.4**
	Iter-G	67.4	48.7	47.6	38.5
	SS-B	64.4	53.8	55.2	53.8
	SS-G	73.0	**56.4**	45.9	43.6

(b) Liberation Day Tariffs

Table 8: Per-query scores across all systems, judge: CLUE. Iter-B/G = Iter-QA-Base/Ginger, SS-B/G = SS-QA-Base/Ginger.

	System	Info F1		Cite F1	
		P	R	P	R
	CAG	65.4	13.2	73.1	2.9
Q1	Bullet	55.6	10.3	66.7	2.9
	Ginger	73.9	13.2	82.6	2.9
	RLM	**88.6**	**29.4**	**94.3**	**17.6**
	Iter-B	0.0	0.0	0.0	0.0
	Iter-G	0.0	0.0	0.0	0.0
	SS-B	13.0	2.9	6.7	1.5
	SS-G	35.7	11.8	3.6	0.0
	CAG	**95.2**	32.4	**100.0**	7.4
Q2	Bullet	89.5	**38.2**	84.2	**23.5**
	Ginger	92.9	32.4	**100.0**	7.4
	RLM	90.6	22.1	90.6	4.4
	Iter-B	0.0	0.0	0.0	0.0
	Iter-G	0.0	0.0	0.0	0.0
	SS-B	8.5	2.9	6.9	1.5
	SS-G	56.4	4.4	0.0	0.0

(a) Nepal Youth Protests

	System	Info F1		Cite F1	
		P	R	P	R
	CAG	**96.9**	8.5	86.7	2.1
Q1	Bullet	88.9	10.1	88.9	1.6
	Ginger	**96.9**	8.5	**90.6**	2.1
	RLM	93.3	16.9	86.7	8.5
	Iter-B	82.4	**18.5**	70.6	4.8
	Iter-G	75.0	12.7	69.4	6.3
	SS-B	56.0	18.0	54.0	**11.6**
	SS-G	58.0	13.8	52.0	5.8
	CAG	97.7	7.9	95.2	2.1
Q2	Bullet	94.4	9.0	**100.0**	3.2
	Ginger	**98.0**	7.9	93.8	2.1
	RLM	90.9	12.2	81.8	5.3
	Iter-B	80.3	13.8	78.7	10.6
	Iter-G	67.4	13.8	52.2	5.3
	SS-B	78.0	**15.3**	71.2	**11.6**
	SS-G	85.1	12.2	68.1	7.4

(b) Palisades Fire

Table 9: Per-query scores across all systems, judge: CLUE. Iter-B/G = Iter-QA-Base/Ginger, SS-B/G = SS-QA-Base/Ginger.

	System	Info F1		Cite F1	
		P	R	P	R
	CAG	62.9	37.9	57.1	12.9
Q1	Bullet	63.6	41.7	66.7	**35.6**
	Ginger	61.1	37.9	58.3	12.9
	RLM	70.0	**46.2**	**70.0**	28.0
	Iter-B	74.6	25.8	3.4	0.0
	Iter-G	**80.8**	24.2	7.7	0.0
	SS-B	17.9	9.8	10.3	3.0
	SS-G	64.8	26.5	57.4	0.0
	CAG	73.0	37.9	70.3	18.9
Q2	Bullet	76.0	**43.2**	84.0	**35.6**
	Ginger	76.5	37.9	70.6	18.9
	RLM	**78.9**	42.4	**92.1**	18.2
	Iter-B	0.0	0.0	0.0	0.0
	Iter-G	0.0	0.0	0.0	0.0
	SS-B	2.4	5.3	9.5	3.8
	SS-G	77.4	15.9	1.6	0.0

(a) Shi Yongxin Scandal

	System	Info F1		Cite F1	
		P	R	P	R
	CAG	79.3	**28.8**	79.3	**23.5**
Q1	Bullet	76.9	**28.8**	80.8	14.2
	Ginger	**85.7**	25.6	**81.6**	18.5
	RLM	**85.7**	22.8	68.6	17.4
	Iter-B	37.7	10.0	15.1	2.1
	Iter-G	29.5	9.3	13.6	3.2
	SS-B	32.3	8.9	29.0	3.9
	SS-G	45.7	7.1	25.0	3.6
	CAG	86.2	15.3	75.9	5.0
Q2	Bullet	84.6	15.7	**84.6**	8.2
	Ginger	**90.3**	15.3	83.3	5.0
	RLM	68.6	**22.4**	62.9	**10.3**
	Iter-B	0.0	0.0	0.0	0.0
	Iter-G	0.0	0.0	0.0	0.0
	SS-B	3.2	6.0	3.2	1.1
	SS-G	11.9	7.5	10.6	1.1

(b) Tropical Storm Wipha

Table 10: Per-query scores across all systems, judge: CLUE. Iter-B/G = Iter-QA-Base/Ginger, SS-B/G = SS-QA-Base/Ginger.

Rank	System	Avg.	nDCG@10	nDCG@20	nDCG@100	R@10	R@20	R@100
2	RRF K=10 + RV	0.759	0.759	0.771	0.811	0.652	0.735	0.832
3	Weighted RRF + RV	0.757	0.757	0.768	0.810	0.650	0.725	0.832
4	RRF K=60 + RV	0.751	0.754	0.765	0.807	0.641	0.716	0.823
5	Sum Sim + RV	0.745	0.747	0.758	0.800	0.636	0.711	0.818
6	RRF K=100 + RV	0.744	0.746	0.757	0.799	0.636	0.711	0.818

Table 11: MAGMaR retrieval final leaderboard positions for MARS submissions. Rank is the public leaderboard rank under the default average over the six displayed retrieval metrics.

Rank	System	Human	Best Votes	Best %	Info P	Info R	Cite P	Cite R
1	ITER-QA-BASE	3.833	8	14.0%	0.347	0.313	0.268	0.258
3	ITER-QA-GINGER	3.694	5	8.8%	0.345	0.290	0.257	0.226
4	SS-QA-GINGER	3.421	10	17.5%	0.544	0.324	0.326	0.238
5	MARQUIS-RLM	3.298	3	5.3%	0.708	0.385	0.592	0.272
6	GINGER	3.123	6	10.5%	0.776	0.404	0.643	0.226
8	SS-QA-BASE	3.070	6	10.5%	0.331	0.306	0.277	0.281
10	BULLET	2.667	0	0.0%	0.711	0.394	0.604	0.237

Table 12: MAGMaR oracle article-generation leaderboard snapshot for MARQUIS submissions. Rank is the public leaderboard rank under the default Human Score ordering.

D.1 General Note Extraction

General note extraction is run independently for each video. The extractor receives the video together with topic and video metadata and produces atomic observations describing directly observable visual content, OCR, and spoken or audio evidence. The output is a JSON object containing a list of notes. Each note includes note text, modality, and an optional timestamp.

D.2 Query-Conditioned Claim Extraction

Query-conditioned claim extraction is run for query–video pairs after aligning each evaluation query to a topic-specific video subset. The extractor receives the query identifier, topic, persona title, background, query text, and video identifier, and outputs claims relevant to the information need. Each claim is tied to a specific query and video and may include confidence, evidence description, source type, and timestamp metadata.

D.3 Query–Topic Alignment

The official query set contains 19 evaluation queries. Each query is aligned to one of 10 topic buckets through deterministic title-to-topic normalization. Query-conditioned claim extraction is then applied over the videos mapped to the corresponding topic. General note extraction uses the topic identity only as metadata and does not condition on the evaluation query.

D.4 Extracted Evidence Schemas

A general note contains a note identifier, video identifier, topic label, note text, modality tag, and optional timestamp. A query-conditioned claim contains a claim identifier, query identifier, video identifier, topic label, claim text, and optional support-oriented metadata such as confidence, evidence description, source type, and timestamp.

D.5 Representative Outputs

To make the extraction flow concrete, we show one representative output from each major stage. These examples are lightly trimmed for presentation but preserve the actual field structure used by the pipeline.

Example general note.

```
{
  "note_id": "gn1a-hol6y3QwX2Y-000",
  "video_id": "hol6y3QwX2Y",
  "topic": "2025_Canadian_Federal_Election",
  "text": "A woman with short blonde hair and
a beige jacket is speaking.",
  "modality": "visual",
  "timestamp": [0.0, 6.0]
}
```

Example query-conditioned claim.

Set	Queries	Total	Min	Avg.	Max
Final flattened retrieval file	19	430	1	22.63	25
Successful decompositions only	18	429	22	23.83	25

Table 13: Sub-query expansion statistics.

2025 Canadian federal election	**2025 Alaska typhoon**
2025 Canadian federal election final seat count by party	2025 Alaska typhoon housing damage assessment report
Elections Canada 2025 election official results dataset	residential structural failures Alaska 2025 typhoon
2025 Canadian federal election popular vote share by party	housing construction materials damaged 2025 Alaska storm
2025 Canadian federal election seat changes by party	geographic distribution of typhoon damage Alaska 2025
2025 Canadian federal election candidate vote totals	roof failure mechanisms 2025 Alaska coastal storm
voter turnout rate Canada 2025 vs 2021	foundation erosion damage coastal Alaska 2025

Table 14: Representative sub-queries produced by the query decomposition stage.


```
{
                                    "claim_id":
"qc-10-1978302738418032640-000",
  "query_id": "10",
  "video_id": "1978302738418032640",
  "topic": "2025_Alaska_Typhoon",
   "claim": "More than 50 people have been
rescued in Western Alaska.",
  "confidence": 0.95,
   "evidence": "Text overlay in the video
states 'More than 50 people have been rescued
in Western Alaska.'",
  "source": "video_text",
  "timestamp": [0.0, 3.0]
}
```


E Appendix: Calibration Implementation

This appendix provides implementation details for video-grounded calibration. Calibration is run after information extraction and assigns a support probability to each extracted artifact without modifying the original artifact content. The calibration prompt itself is provided in Appendix I.

E.1 Calibration Inputs and Outputs

For each extracted artifact, the calibration stage receives the source video and the artifact text. The output is a scalar support probability in $[0, 1]$ estimating whether the artifact is supported by the source video. The calibrated artifact preserves the original note or claim and attaches a calibration payload containing the support score and backend provenance.

E.2 Backends

We evaluate two calibration backends. The prompted backend uses Qwen3.5 with a constrained probability-estimation prompt. The comparison backend is CLUE built on the Qwen2.5-Omni family. Both backends consume the same video–artifact pairs and emit the same conceptual output type.

E.3 Attachment Logic

Calibration predictions are attached to artifacts using stable artifact identifiers when available. When identifiers are unavailable or inconsistent, the attachment stage falls back to matching by video identifier and artifact text. This preserves compatibility across extraction and calibration jobs while keeping the original artifact representation unchanged.

Example calibrated artifact.


```
{
                                    "claim_id":
"qc-10-1978302738418032640-000",
   "claim": "More than 50 people have been
rescued in Western Alaska.",
  "calibration": {
    "unli": {
      "prob": 0.95,
      "raw": {
```

Sub-queries kept	nDCG@10	nDCG@100	R@10	R@100
1 random	0.613 ± 0.029	0.679 ± 0.025	0.512 ± 0.027	0.707 ± 0.024
5 random	0.684 ± 0.034	0.749 ± 0.029	0.601 ± 0.041	0.794 ± 0.031
10 random	0.696 ± 0.010	0.763 ± 0.013	0.615 ± 0.018	0.812 ± 0.015
All	0.711	0.773	0.640	0.831

Table 15: Effect of randomly retaining fewer sub-queries before max-similarity fusion. Random conditions are averaged over five seeds.

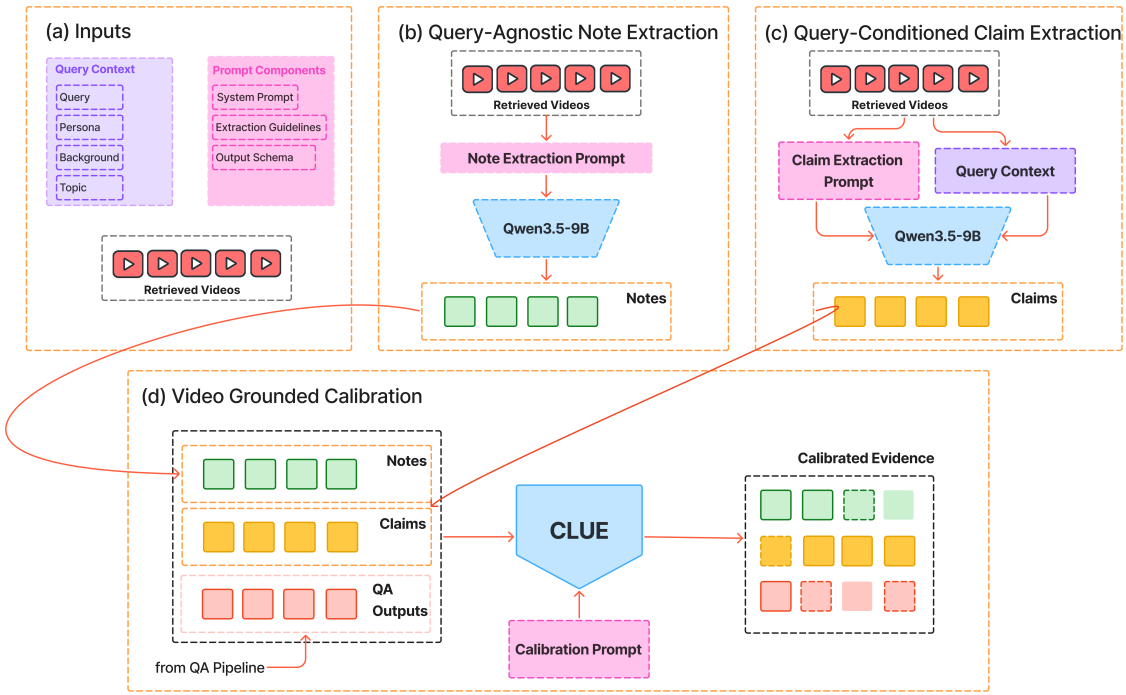


Figure 2: Information extraction and calibration workflow. Retrieved videos and prompt components are used to produce query-agnostic notes and query-conditioned claims, while QA outputs enter from the question-answering pipeline (See Appendix F and Figure 3). These extracted evidence records are merged and scored against source video by the calibration backend, producing calibrated extracted evidence for article generation.

```
        "raw_output": "<answer>0.95 </answer>"
      }
    }
  }
}
```

E.4 Claim Filtering

In addition to attaching support probabilities, the calibration stage can optionally filter extracted claims using the predicted support score. Claims with support probabilities below a predefined threshold are excluded from downstream outputs. This filtering mechanism is intended to reduce unsupported or weakly grounded claims while preserving high-confidence artifacts.

The filtering threshold is configurable and applied uniformly across calibration backends. Importantly, filtering is performed only after extraction and does not modify the original extracted content or calibration predictions.

For the calibration models(CLUE and Qwen3.5), the hyperparameters are summarized in Table 4.

F Appendix: Question Answering Implementation

Figure 3 provides an overview of the single shot and iterative question answering systems.

Single-Shot Question Answering. The single-shot pipeline uses Qwen3.5-27B for answer generation, Qwen2.5-Omni-7B with OmniEmbed-

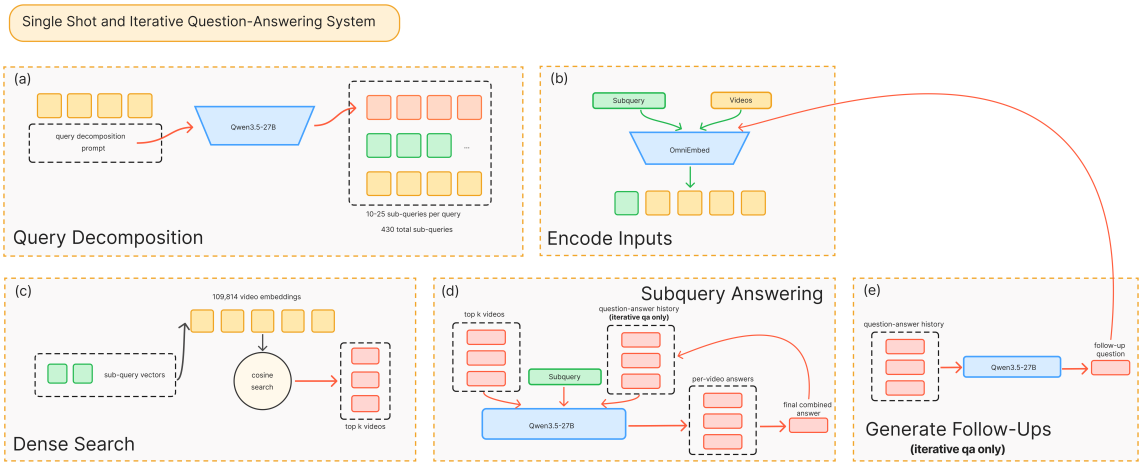


Figure 3: Overview of QA-based evidence extraction method. The single-shot variant decomposes the query into fixed subqueries, retrieves videos, answers each subquery, and aggregates the answers. The iterative variant generates follow-up questions from the question-answer history until a stopping condition is met.

v0.1 for multimodal embeddings, and Whisper (medium.en) for transcription.

Videos are preprocessed by downsampling to 30 FPS with a max frames of 32. Transcripts are generated from the audio stream and paired with each video. Video and query embeddings are computed in a shared space using the OmniEmbed model, and retrieval is performed using cosine similarity with a threshold of 0.1 and top-k selection with k=4.

For each sub-query, the system retrieves relevant videos and generates per-video answers using Qwen3.5-27B, conditioned jointly on the video frames, transcript, and query. The model is prompted to produce concise factual answers grounded only in the provided inputs. All per-video answers are collected, and responses such as "I don't know" are filtered out. The remaining answers are then merged using a second Qwen3.5-27B pass that combines the extracted answers into a single response without introducing external knowledge.

Across the evaluated queries, the single-shot pipeline generated on average 23.84 expanded questions per main query, with a minimum of 20 and a maximum of 25 expanded questions. In the single-shot setting, 246 of 453 expanded questions were unanswered. Qualitative examples from the single-shot output for the first main query are shown below.

Expanded question	Answer
What were the final seat counts for each political party in the 2025 Canadian federal election?	Based on the provided video transcript with 96% of votes counted, the projected seat counts were Liberals: 166, Conservatives: 146, Bloc Québécois: 23, and NDP: 7. The final official counts after 100% of votes were counted were unknown based on the provided text.
What is the official dataset published by Elections Canada containing detailed results for the 2025 federal election?	I don't know. The specific name of the official dataset published by Elections Canada containing detailed results for the 2025 federal election is unknown based on the provided information.
What were the popular vote percentages for the Liberal Party and Conservative Party in the 2025 Canadian federal election?	The Liberal Party received 43.1% of the popular vote and the Conservative Party received 41.7%.

Table 16: Examples of single-shot expanded questions and answers for the first main query.

Iterative Question Generation. The iterative pipeline uses the same underlying structure as the single-shot approach, so each step performs retrieval and per-video answer generation in the same way as the single-shot pipeline.

The key difference is that instead of processing each sub-query once, the system maintains a running history of question–answer pairs and iteratively refines the query. After each retrieval and answer aggregation step, the aggregated answer is appended to the history, and a new follow-up question is generated using Qwen3.5-27B with sam-

pling enabled, prompting the model to produce exactly one question that extracts additional or more specific information conditioned on the full history. This loop continues for up to 5 steps per sub-query but terminates early if no videos are retrieved, the model outputs "NONE" as the next question, or a repeated question is detected.

Across the evaluated queries, the iterative pipeline generated a minimum of 22 expanded questions, a maximum of 73 expanded questions, and an average of 41.05 expanded questions per main query. In the iterative setting, 293 of 613 expanded questions were unanswered.

Qualitative examples from the iterative output for the first main query are shown below. These examples illustrate how the iterative method can generate follow-up questions that become more specific than the original expanded questions.

Expanded question	Answer
What were the final seat counts for each political party in the 2025 Canadian federal election?	Based on the provided transcript with 96% of votes counted, the projected seat counts were Liberals: 166, Conservatives: 146, Bloc Québécois: 23, and NDP: 7. The final official seat counts were unknown because the provided data represented projections before 100% of votes were counted.
Which specific electoral districts accounted for the largest swing in votes that resulted in the projected reduction of the NDP's seat count to seven?	Fortress Vancouver, Fortress Montreal, and the GTA accounted for the largest swing in votes that resulted in the projected reduction of the NDP's seat count to seven.

Table 17: Examples of iterative expanded questions and answers for the first main query.

G Appendix: Article Generation Implementation

This appendix provides implementation details for the article generation systems. Prompt templates for all article generation variants are provided in Appendix I.

G.1 Evidence Inputs

The article generation systems operate over flat lists of evidence artifacts. These artifacts may be query-conditioned claims, query-agnostic notes, or QA pairs. Claims and notes include video identifiers and timestamps when available. QA pairs include the source videos used to produce the answer.

G.2 BULLET Generation

The bullet-point generator renders selected evidence items directly as a numbered list of findings with inline citations. This variant does not invoke a generation model and is intended as a conservative evidence-presentation baseline.

G.3 Single-Prompt Article Generation

The single-prompt article generator concatenates the evidence items for a query into a single prompt and generates a coherent article with inline citations. To reduce context length and memory failures, evidence sets larger than 25 items are truncated to the top 25 by confidence score.

G.4 GINGER Article Generation

The GINGER generator decomposes article generation into facet clustering, cluster ranking, per-cluster summarization, and fluency enhancement. Since the information extraction stage already produces atomic evidence units, the pipeline begins from extracted notes, claims, or QA pairs rather than running a separate nugget-detection stage.

H Appendix: RLM Controller Implementation

This appendix documents the tool API, memory schema, and prompts used by the RLM controller (see section 6 in the main text). The goal is to make the RLM-side of our submission directly reproducible. Figure 4 summarizes the resulting Think–Act–Observe control loop.

H.1 Tool API and Backing Modules

Table 18 lists all callable tool functions registered in the REPL namespace.

H.2 Memory Bank JSON Structure

```
{
  "findings":      [" high-level
     insights "],
  "keywords":      {"<video_id>": ["
     keyword1", "keyword2"]},
  "fact_table":    {"<video_id>": [
    {"fact": "...", "timestamp": "10s
       -15s", "source_tool": "
       query_claims", "confidence":
       0.8}
  ]},
  "selected_facts": ["facts chosen by
      llm_judge for the final report
     "],
  "videos":        {"<video_id>": {"
     status": "processed", "
```

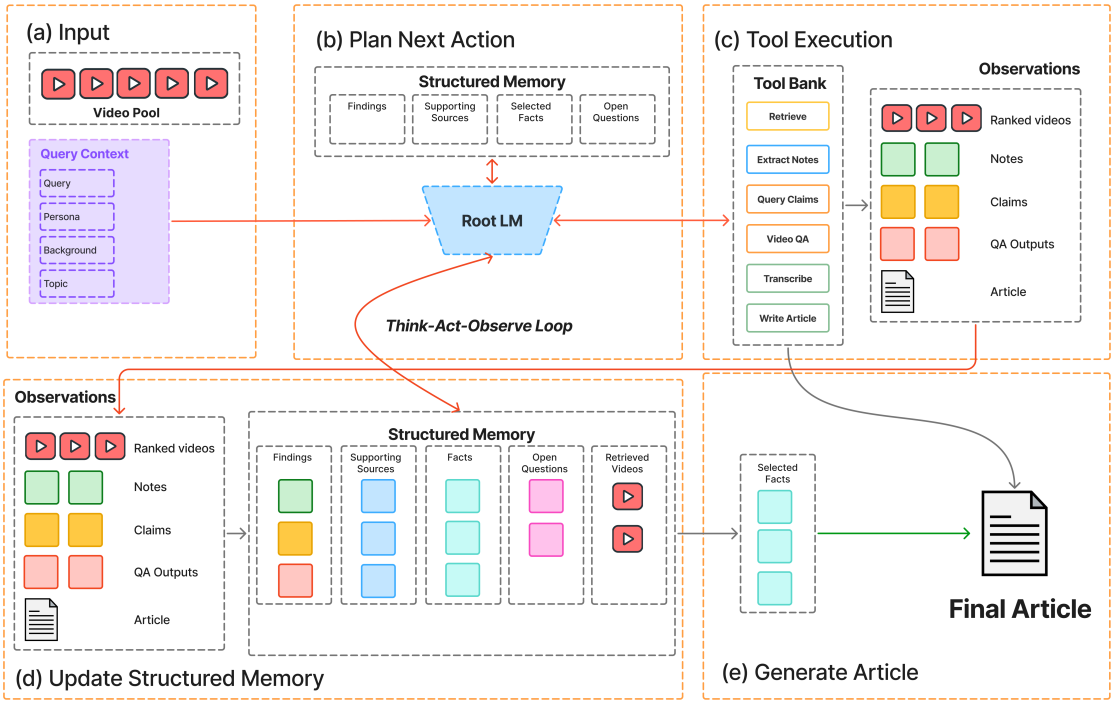


Figure 4: MARQUIS-RLM controller. The Root LM reads structured memory, plans the next action, executes one tool call in a persistent REPL environment, observes the result, and updates memory before continuing. Once sufficient evidence has been gathered and judged, selected facts are passed to the article-generation tool to produce the final cited article.

```
        tools_used": [...], "path": "
        ...", "caption": "..."}}
}
```

H.3 Examples of Root LM Behavior

Example 1 Final recheck.

```
[THINK] (6.5s):
  The report has been generated but gaps remain. I
should verify one final critical issue - whether the
seat counts (Lib 120, CON 87) shown in zaFtBz84Kyk
were final or still live projections, since this
affects report accuracy. Let me verify this with
video_qa one more time.

[ACT]:
  ```repl
  video_qa("zaFtBz84Kyk",
           "Was the seat count shown (Lib 120, CON 87)
 the final result or a live projection? Clarify the
exact time shown and when results were finalized.")
  ```

In [42]:
  1 video_qa("zaFtBz84Kyk", "Was the seat count shown
 (Lib 120,
              CON 87) the final result or a live
projection? ...")

Out [42]:
  '[video_qa] zaFtBz84Kyk - Q: Was the seat count
shown (Lib 120, CON 87) the final result or a live
projection? Clarify the exact time shown and when
results were finalized. A: The seat count shown (Lib
120, CON 87) was a live projection, not the final
result, as the final declaration occurred later with
the banner "LIB WIN" and the count updated to 124 for
Liberals. confidence=1.0, timestamp=220-230s'

[OBSERVE]:
  TOOL CALLS (1):
    OK video_qa({'video_path': 'zaFtBz84Kyk', ...})
-> 82.7s

[JUDGE] iter 41 (model: qwen/qwen3.5-9b):
  Reasoning 5 | Action 5 | Granularity 5 | Progress 5
  Redundancy 5 | Think 5 | Code 5 | Output 5 | Tool 5
  WEIGHTED TOTAL: 62.5/62.5 (100.0%)
```

Observation. Root LM re-engages `video_qa` to disambiguate a data point to prioritize correctness over completion.

Example 2 Vague Information Clarification.

```
[THINK] (3.4s):
  I've watched all 17 videos but only queried 4 for
facts. I'm missing critical casualty and aid
distribution data. Video 1978302738418032640 caption
mentions "at least one is dead and several remain
missing" - I need to extract these numbers via
video_qa.

[ACT]:
  ```repl
  video_qa("1978302738418032640","How many fatalities
and missing persons are there?What details about
```
```

Tool	Signature	Backing module
Perception tools (multimodal extraction over raw video)		
Caption	*video_caption(vid)*	local Qwen3.5-9B model
GeneralNotes	*general_notes(vid)*	subsection 4.1 general note extraction
QueryClaims	*query_claims(vid)*	subsection 4.2 query-conditioned claim extraction
Targeted-query tools		
VideoQA	*video_qa(vid, question)*	subsection 4.3 multimodal QA pipeline
Transcribe	*transcribe(vid)*	local Whisper Model
RetrieveChunks	*retrieve_chunks(vid)*	section 3 OmniEmbed retriever, lowered to 20 s chunk level
Generation tool		
WriteReport	*write_report(facts)*	section 5 GINGER-based pipeline
Memory operators		
memory_summary	*memory_summary()*	compact memory snapshot per iteration
print_memory	*print_memory(slot=None)*	full JSON dump of one or all slots
add_keyword	*add_keyword(vid, kw)*	tag a video with a keyword
search_by_keyword	*search_by_keyword(kw)*	find items in memory bank
remove_fact	*remove_fact(vid, idx)*	delete a single fact
clear_facts	*clear_facts(vid=None)*	clear all facts for one or all videos
Memory operators		
Think	*llm_think()*	fact_table → findings (infered by LLM)
Judge	*llm_judge()*	fact_table → selected_facts (for report generation)

Table 18: Tools and memory operators registered in the RLM REPL namespace.

```
rescue operations?")
  ```

In [27]:
  1 video_qa("1978302738418032640", "How many
fatalities and missing persons are there? ...")

Out [27]:
  '[video_qa] 1978302738418032640 - Q: How many
fatalities and missing persons are there? What details
 about rescue operations?
   A: At least one fatality and several missing; over
50 rescued via helicopter.
confidence=1.0, timestamp=8-15s'

[OBSERVE]:
  TOOL CALLS (1):
    OK video_qa({'video_path': '1978302738418032640',
 ...}) -> 22.2s

[JUDGE] iter 26 (model: qwen/qwen3.5-9b):
  Reasoning 5 | Action 5 | Granularity 5 | Progress 5
  Redundancy 5 | Think 5 | Code 5 | Output 5 | Tool 5
  WEIGHTED TOTAL: 62.5/62.5 (100.0\%)
```

**Observation.** Root LM identified the information from the caption was vague and decide to call the QA tool to get a more precise answer.

**Example 3 Cross-modal Conflict Resolution.**

```
[THINK] (4.1s):
  I need to resolve the major conflict - seat counts
vary between r2P6c5tzvM8 (LIB 152, CON 120) and
zaFtBz84Kyk (LIB 119-120, CON 84-87). r2P6c5tzvM8
appears more authoritative as CTV News official report
. I should use transcribe to get the final declared
results from both CTV videos to reconcile this
conflict.

[ACT]:
  ```repl
  transcribe("r2P6c5tzvM8")
  ```
```

```
In [17]:
  1  transcribe("r2P6c5tzvM8")

Out [17]:
  "Transcript of r2P6c5tzvM8 (2325 chars):
   If you're Canadian, keep watching. It's election
time. I've just requested that the Governor General
dissolve Parliament and call it an election, and she
has agreed. The campaign will last 36 days, and the
election will be on April 28th. There are 343 ridings
up for grabs across the country. ... Here's what the
House looks like right now with the Liberal minority
government. The Liberals have 152 seats. The
Conservatives have 120. The Bloc Quebecois have 33.
The NDP have 24. The Green Party has 2. Three MPs are
sitting as independents, and four seats are vacant.
For a majority government, a party needs to win at
least 172 seats, but a minority government can be
formed with fewer than that. ..."

[OBSERVE]:
  TOOL CALLS (1):
    OK  transcribe({'video_path': 'r2P6c5tzvM8'})  ->
 32.2s

[JUDGE] iter 16 (model: qwen/qwen3.5-9b):
  Reasoning 4 | Action 5 | Granularity 5 | Progress 4
  Redundancy 5 | Think 5 | Code 5 | Output 5 | Tool 5
  WEIGHTED TOTAL: 60.5/62.5 (96.8%)
```

Observation. When root gets conflicting visual information, it decides to call `transcribe` to use transcript evidence for cross-modal conflict resolution.

H.4 Statistics

Tool Usage. As shown in Table 19, average tool usage indicates a strongly adaptive exploration strategy. In practice, the agent schedules all tools in 14 of 19 queries. In practice, it uses Caption tool (10 ± 3) mainly as a relevance filter, and relies more
```

	Tool Calls (Average)						
	Capt.	Claims	Notes	QA	Trans.	Retr.	Total
Q1	6	6	4	6	4	1	28
Q2	6	6	4	6	2	2	26
Q3	5	6	5	4	2	2	24
Q4	5	5	2	2	2	0	15
Q5	8	6	2	2	1	1	21
Q6	7	10	4	4	2	2	29
Q7	11	12	3	4	1	1	32
Q8	10	10	4	2	2	1	28
Q9	7	7	1	1	0	0	16
Q10	17	5	3	4	0	1	30
Q11	18	10	2	2	0	0	32
Q12	11	8	7	2	3	2	33
Q13	10	5	6	2	2	0	24
Q14	10	7	3	2	1	1	24
Q15	10	8	4	3	2	1	28
Q16	10	10	4	1	1	1	26
Q17	11	12	6	2	1	0	32
Q18	10	10	2	1	1	2	26
Q19	10	10	1	1	1	1	24
Avg.	10±3	8±2	4±2	3±2	2±1	1±1	26±5

Table 19: Statistics of tool calls per query (averaged over two runs).

on targeted extraction with QueryClaim tool(8 ± 2) than on broader summarization with General Notes tool (4 ± 2). Transcribe tool is used only as a supplement, while Retrieval tool is rarely needed (1 ± 1), since most videos are short (< 60s) and do not require fine-grained temporal localization.

Runtime Performance. The evaluation by LLM-as-a-judge indicates strong behavioral quality (92±2% on average), with near-perfect Output Waste and Code Minimality scores as reported in Table 20. The average wall time per query is 36±12 minutes, with most of the runtime spent on tool execution rather than Root LM inference.

Root LM Context Window Usage As detailed in Table 21, Root LM use an average of only 33% of its 32K context window, avoiding frequent truncation. The context growth scales linearly to roughly 1.3% per iteration (from 8% at Iteration 0 to 61% at Iteration 40). This stable progression indicates that the accumulation of both the Memory Bank and history is controllable, therefore preventing explosive token consumption in later iterations.

I Appendix: Prompt Templates

This appendix collects all prompt templates used by the system. Prompts are grouped by the method component that uses them. Implementation details for each method are provided in Appendices C–H.

I.1 Retrieval Query Expansion Prompt

Sub-query decomposition prompt. The decomposition prompt consumes the structured fields of a MAGMaR query and emits a JSON array of fine-grained search phrases. It is used with Qwen3.5-27B and thinking disabled.

```
You are a research decomposition specialist.
Your task is to take a user's query
and break it down into an exhaustive set of
searchable sub-queries - short
phrases or keyword combinations that could be
entered into a search engine or
database to retrieve all the information
needed to fully answer the original query.
You will receive the following inputs:
- Title: <TITLE>
- Language: <LANGUAGE>
- Persona: <PERSONA_TITLE>
- Background: <BACKGROUND>
- Query: <QUERY>
Decomposition Rules:
1. Coverage: Extract every distinct piece
of information the user is asking for. Do
not merge separate information needs into one
sub-query.
2. Granularity: Each sub-query should
target ONE specific, retrievable piece of
information. Prefer atomic queries over
compound ones.
3. Implicit needs: Go beyond what is
explicitly stated. Based on the background
and persona_title, infer what additional
information the user would likely need but
did not explicitly ask for.
4. Search-friendly format: Each sub-query
should be phrased as a concise search phrase,
typically 3-10 words, not a full sentence or
question.
5. Context anchoring: Each sub-query should
include enough context to be independently
searchable without ambiguity.
6. Source-awareness: If the user requests
source information, generate sub-queries
targeting official sources, methodologies,
and data provenance.
7. Dimensional expansion: Consider additional
perspectives or breakdowns by time, place,
category, cause, mechanism, or comparison only
when they add value.
8. No redundancy: Each sub-query must be
meaningfully distinct.
9. Language: Always generate sub-queries in
English.
10. Generate between 10 and 25 sub-queries.
11. Do not mechanically prepend the full topic
title to every sub-query.
12. Focus on the specific information being
sought, not on repeating the topic name.
Return ONLY a JSON array of strings. No
explanation, no markdown, no code blocks.
```

General note extraction prompt. The general-note prompt is query-agnostic but not fully context-free: it includes the source topic and video iden-

	Quality		Latency		
	Judge (%)	Facts/Iter	Wall (min)	Tokens (K)	Root LLM (s)
Q1	88	0.74	46	595	382
Q2	92	0.57	47	510	326
Q3	86	0.90	30	516	311
Q4	94	1.85	24	329	216
Q5	95	1.11	36	405	278
Q6	92	0.37	49	476	342
Q7	92	1.08	42	493	416
Q8	89	1.72	48	563	409
Q9	91	1.39	16	175	199
Q10	93	0.99	26	418	357
Q11	93	1.56	22	430	300
Q12	93	1.28	22	472	481
Q13	92	1.00	22	271	346
Q14	92	0.93	36	457	366
Q15	89	1.21	45	470	351
Q16	92	0.86	48	613	342
Q17	90	0.81	54	419	298
Q18	94	0.91	37	378	235
Q19	95	1.26	26	329	244
Avg.	92±2	1.1±0.4	36±12	438±109	326±72

Table 20: Statistics of behavior quality and query latency, averaged over two runs.

Metric	Value	
Per-Iteration Token Consumption		
Total tokens	11131	
Prompt (context)	10689 ± 6434	(96%)
Completion	442 ± 415	(4%)
Reasoning	328	
Output	114	
Context Window Utilization (32K limit)		
Average usage	33% ± 20%	
Maximum usage	97%	
Iterations >80%	25	(1.7%)
Context Growth Over Iterations		
Iteration 0	2,582	(8%)
Iteration 20	9,983	(31%)
Iteration 40	20,026	(61%)

Table 21: Root LM context-window usage, averaged over two runs.

tifier together with an evidence-first instruction block.

```
You are extracting observation notes directly
from a raw video.
Video context:
- topic: <TOPIC>
- video_id: <VIDEO_ID>
- timestamp_span: <TIMESTAMP_SPAN_OR_NULL>
Rules:
- Record only directly observable content.
- No inference, speculation, causality, or
cross-video synthesis.
- Capture OCR (on-screen text), events, and
visible scene details.
- One note per atomic visible, audible, or
textual fact.
- Use modality 'visual' for scene content,
'ocr' for on-screen text, and 'audio' for
transcript or speech.
- Use the provided timestamp span for each note
when no narrower timestamp is available.
- If there is no usable evidence, return an
empty notes list.
Output strict JSON only.
No markdown, no code fences, no explanation,
no extra keys outside the schema.
Expected shape:
{
  "notes": [
    {
      "text": "...",
      "modality": "visual",
      "timestamp": [0.0, 1.0]
    }
  ]
}
```

Query-conditioned claim extraction prompt. The single-query claim-extraction prompt conditions on the query text together with persona, background, topic, and video identity.

```
You are extracting query-relevant claims
directly from a raw video.
Query context:
- query_id: <QUERY_ID>
- topic: <TOPIC>
- persona_title: <PERSONA_TITLE>
- background: <BACKGROUND>
- query: <QUERY_TEXT>
- video_id: <VIDEO_ID>
Rules:
```

```
- Extract up to <PER_VIDEO_TARGET> claims relevant to the query from this video.
- Claims must be directly supported by observable video content.
- Avoid generic scene summary unless it directly serves the query.
- Avoid duplicates and paraphrases.
- If the video does not contain evidence for the query, return an empty claims list.
- source must be one of 'video_visual', 'video_text', or 'transcript'.
- timestamp must be [start, end].
- confidence must be a float between 0 and 1.
Output strict JSON only.
No markdown, no code fences, no explanation, no extra keys outside the schema.
Expected shape:
{
  "claims": [
    {
      "claim": "...",
      "confidence": 0.85,
      "evidence": "...",
      "source": "video_visual",
      "timestamp": [0.0, 1.0]
    }
  ]
}
```

Decompose Query into Questions. Expand single query into a series of question–answer pairs based on a provided title, language, persona, and background.

```
You are a research decomposition specialist. Your task is to take a user's query and break it down into an exhaustive set of searchable research questions — complete questions that could be used to retrieve all the information needed to fully answer the original query.
You will receive the following inputs:
- Title: {title}
- Language: {language}
- Persona: {persona_title}
- Background: {background}
- Query: {query}
Decomposition Rules:
1. Coverage: Extract every distinct piece of information the user is asking for. Do not merge separate information needs into one question. If the query asks for multiple related but distinct data points, each one should become its own question.
2. Granularity: Each question should target ONE specific, retrievable piece of information. Prefer atomic questions over compound ones.
3. Implicit needs: Go beyond what is explicitly stated. Based on the background and persona_title, infer what additional information the user would likely need but did not explicitly ask for. Consider what a professional in that role would typically require to produce complete, high-quality work on this topic.
4. Search-friendly format: Each sub-query must be written as a concise, well-formed question that could plausibly be entered into a search engine or research database.
5. Context anchoring: Each question should include enough context (e.g., specific names, dates, locations, technical terms) to be independently searchable without ambiguity.
6. Source-awareness: If the user requests source information or credibility indicators, generate questions specifically targeting official sources, methodologies, and data provenance.
7. Dimensional expansion: For each core information need identified, consider whether the user would benefit from additional perspectives or breakdowns. Ask yourself: can this information be meaningfully decomposed further by time, place, category, cause, mechanism, comparison, or any other axis that is natural and relevant to the topic? Only expand along dimensions that genuinely add value given the query's subject matter and the user's background.
8. No redundancy: Each question must be meaningfully distinct. Do not produce near-duplicates that would return the same search results.
9. Language: Always generate questions in English, regardless of the language field in the input.
10. Quantity: Generate between 10 and 25 questions. Focus on quality and relevance over quantity.
11. Avoid mechanical repetition: Do not mechanically prepend the full topic title to every question. Each question should contain only the context necessary for an effective search.
12. Focus on information needs: Focus on the specific information being sought rather than repeating the topic name unnecessarily.
Return ONLY a JSON array of strings. No explanation, no markdown, and no code blocks.
For example, given a query about the 2025 Canadian federal election asking for seat counts and vote shares, good questions would be:
[ "What was the total number of seats won by each political party in the 2025 Canadian federal election?", "What percentage of the national popular vote did each major party receive in the 2025 Canadian federal election?", "How many seats did each party gain or lose compared with the 2021 Canadian federal election?", "What official datasets published by Elections Canada contain vote totals and seat counts for the 2025 federal election?", "What demographic voting patterns were observed in the 2025 Canadian federal election?" ]
NOT:
[ "What happened in the 2025 Canadian federal election?", "What were the results of the 2025 Canadian federal election?", "What information is available about the 2025 Canadian federal election?" ]
JSON array:
```

Question Answering. Qwen 3.5 produces an answer to the question based on the down-sampled video and transcript.

```
Question: question
Answer concisely using the video and transcript. Answer to the best of your abilities. If you can't answer all of the question, answer the parts that you can (Ex. if asked about Liberal and Conservative vote counts, but only have the Liberal vote counts, return those). If you have no information about anything related to the question, return "I don't know". Never say the event hasn't taken place.
Return ONLY the final factual answer.
```

Combined Answers. Combines all the answers generated independently from each relevant video in the top k most similar.

```
You are given extracted answers from videos. These answers are factual and must be used.
Question: {subquery}
Extracted Answers (treat as ground truth): {valid_answers}
Combine them into a single answer.
Do NOT use prior knowledge. Do NOT say the event has not taken place. Only use the provided answers.
If you recieve conflicting information, make a best guess NOT on prior knowlege but based on the nature of the question (Ex. for a question about how many seats a party has one, its reasonable to assume the largest number is the most recent seat count)
Return only the final answer. You might not be able to answer it fully, and that's okay. Answer what you can and then say specifically what information is unknown.
```

Generate Follow Up Question. Generates the follow up question in the iterative QA system based on entire past context.

```
You are refining a research question based on prior answers.
Context: context
Generate ONE new question that: - extracts new information not yet covered - is more specific or differently framed
If no meaningful new question can be formed, output: NONE
Return ONLY the question or NONE.
```

Qwen 3.5 calibration prompt. The prompted Qwen 3.5 backend receives the full source video together with one extracted artifact and is instructed to return a scalar probability in a constrained answer format.

```
To help you make more accurate and consistent judgments, here is an expanded explanation of how to interpret and assign support percentages.
These examples are designed to cover a range of real-world cases you may encounter in the annotation task.
100% - /Full and unambiguous support:
The video clearly shows the exact event described in the claim. There is no need for guessing or interpretation.
80-100% - Almost complete support:
The main content in the claim is shown, but there may be minor ambiguity in location, identity, or completeness. The overall claims are supported by the video.
60-80% - Strong partial support:
The video strongly suggests the claim is true, but some critical details may be missing, obscured, or ambiguous, limiting the ability to confirm the claim with certainty. The video gives strong but not definitive support.
40-60% - Moderate partial support:
There is some alignment with the claim, but large portions are either missing, unclear, or open to interpretation. While the footage may point in the same general direction as the claim, it lacks the clarity or completeness needed for confident verification.
20-40% - Minimal weak support:
There are small visual or audio cues that could hint at the claim, but they are insufficient to be confident.
0-20% - Very weak or speculative support:
There may be the slightest indirect reference, such as a related object or setting, but nothing concrete happens.
0% - No support or contradiction:
The video does not relate to the claim at all, or it directly shows something opposite.
Based on the provided video and text, evaluate the probability that the text is true.
Your answer must be a decimal number between 0 and 1, and you must strictly follow the format below:
<answer>probability_value</answer>
Where probability_value is the result you calculate.
The text to evaluate is:
<ARTIFACT_TEXT>
```

Malformed outputs trigger a stricter retry prompt that preserves the same task content while requiring a single answer in the exact form `<answer>0.73</answer>`.

Baseline The baseline approach feeds all query-conditioned claims into a single LLM prompt and generates the complete report in one forward pass.

```
You are a report writing assistant. Your task is to synthesize a set of claims extracted from multiple videos into a single, fluent, well-organized report that answers the given query.
```

```
## Instructions:
1. Read all the claims below carefully. Each claim was extracted from a specific video and has a timestamp.
2. Group related claims together logically (e.g., by sub-topic or chronological order).
3. Write a coherent, well-structured report that covers all the key information from the claims.
4. For EVERY piece of information in your report, include an inline citation in the format [video_id, timestamp_start-timestamp_end].
5. If multiple claims from different videos support the same point, cite all relevant sources.
6. Remove redundant information — if multiple claims say the same thing, mention it once and cite all sources.
7. The report should be fluent and readable, not a list of bullet points.
8. Keep the report concise but comprehensive (aim for 200-400 words).
## Query/Topic: {topic}
## Claims:
{claims_text}
## Report:
```

GINGER clustering prompt. The model receives all claims for a query and is instructed to partition them into thematic facet clusters, returning a labeled JSON partition of the claim set.

```
You are an information analyst. Given a set of claims about a topic extracted from videos, group them into distinct facet clusters. Each cluster should represent a different sub-topic or aspect of the main topic.
## Instructions:
1. Read all claims carefully.
2. Group them into clusters based on their sub-topic/facet (e.g., "casualties", "rescue efforts", "damage assessment", "government response", etc.).
3. Each claim should belong to exactly one cluster.
4. Give each cluster a short, descriptive label.
5. Output your result as a JSON object with the following format:
{
"clusters": [
{
"label": "Short descriptive label for this facet",
"claim_ids": ["qc-10-xxx-000", "qc-10-xxx-001"]
},
...
]
}
Only output the JSON object, no other text.
## Topic: {topic}
## Claims:
{claims_text}
```

GINGER ranking prompt. The model receives the labeled clusters and is instructed to rank them by relevance to the query topic, returning an ordered JSON array of cluster labels.

```
You are a relevance assessor. Given a query/topic and a list of facet clusters (each containing grouped claims from videos), rank the clusters from most to least relevant to the query.
## Instructions:
1. Consider which facets are most important for answering/addressing the query topic.
2. Rank all clusters from most relevant to least relevant.
3. Output a JSON array of cluster labels in order from most to least relevant:
{
"ranked_labels": ["most relevant label", "second most relevant", ...]
}
Only output the JSON object, no other text.
## Topic: {topic}
## Clusters:
{clusters_text}
```

GINGER summarization prompt. The model receives the claims within a single cluster and is instructed to condense them into one cited sentence of at most 40 words, preserving inline citations anchored to the supporting evidence.

```
You are a concise summarizer. Summarize the following cluster of claims into a SINGLE sentence (maximum 40 words). The sentence must:
1. Capture the key information from all claims in this cluster.
2. Include inline citations in the format [video_id, timestamp] for every fact mentioned.
3. Be factual — only include information present in the claims.
## Cluster: {cluster_label}
## Claims in this cluster:
{cluster_claims_text}
## One-sentence summary:
```

GINGER fluency prompt. The model receives the concatenated one-sentence cluster summaries and is instructed to rewrite them into a coherent 200–400-word prose report without adding new information or removing any citations.

```
You are an editor. Below is a report composed of individual summary sentences about the topic "{topic}". Your task is to rewrite it into a smooth, fluent, well-organized report.
## Rules:
1. Do NOT add any new information that is not in the summaries below.
```

```
2. Do NOT remove any information or citations
from the summaries.
3. Keep ALL inline citations in the format
[video_id, timestamp].
4. Improve transitions between sentences for
better readability.
5. You may reorder sentences for better
logical flow.
6. Keep the report concise (200-400 words).
## Draft report (concatenated summaries):
{draft_report}
## Final polished report:
```

MARQUIS-RLM REPL system prompt.

```
You answer queries using an interactive Python
REPL, called iteratively until you submit a
final answer.
THINK-ACT-OBSERVE LOOP:
Each iteration: THINK (brief reasoning), ACT
(one code block), OBSERVE the output.
THINK phase: READ the memory snapshot below
— it shows your findings (global knowledge)
and per-video facts. Base your next action on
what you ALREADY KNOW, not assumptions.
{pacing}
ENVIRONMENT:
- context['task'], context['video_pool'],
context['tools'] are read-only.
- ‘memory‘ is a persistent dict (survives
compaction).
- Tools are pre-loaded as plain Python
functions; call them directly.
FORMAT: THINK (2-4 sentences), then ONE
“‘repl“‘ code block (1-5 lines, ONE tool call).
NO for-loops over videos.
FINAL ANSWER: report =
write_report(memory['selected_facts']),
then FINAL_VAR(report) outside the code
block.
```

MARQUIS-RLM Root LM Think prompt.

```
TASK: {query_text}
CURRENT FINDINGS:
{findings_str}
FACT TABLE SUMMARY:
{fact_summary}
VIDEO STATUS:
{video_status}
You are the analytical brain. Based on all
facts collected so far:
1. NEW_FINDINGS: List any new high-level
findings (one sentence each) not already in
CURRENT FINDINGS. If a new fact CONTRADICTS
an existing finding, say CONFLICT: <existing>
vs <new>.
2. UPDATED_FINDINGS: Output the complete
updated findings list (old + new,
deduplicated). One finding per line,
prefixed with ‘- ’.
3. NEXT_STEPS: What should the agent do next?
Be specific: which video, which tool, which
question.
Be concise.
```

MARQUIS-RLM Root LM Judge prompt.

```
TASK: {query_text}
FINDINGS (root's current understanding):
{findings_str}
FACT TABLE ({n} facts):
{fact_lines}
You are a strict quality judge. Review ALL
facts above for the task.
1. ITEM REVIEW: For each fact (F#0, F#1, ...),
give a verdict.
BE CONSERVATIVE — only REMOVE if clearly
irrelevant or duplicate. When in doubt, KEEP.
KEEP — useful, specific, or even mildly
relevant (default)
REMOVE — clearly irrelevant or duplicate of
another listed fact
REWRITE — needs more detail or has a missing
timestamp (flag, do NOT drop)
Format: F#0: KEEP / F#3: REMOVE (dup of F#1)
/ F#5: REWRITE (missing timestamp)
2. SELECTED: Pick the 10-40 BEST facts for a
comprehensive report (prefer MORE coverage).
List their IDs: SELECTED: F#0, F#2, F#7, ...
3. MISSING TIMESTAMPS: List facts that are
useful but lack timestamps; suggest video_qa
queries to resolve them.
4. GAPS: What information is still missing for
a thorough report?
5. READY: Can we write a good report now? (yes
/ no / almost)
Be specific and concise.
```

MARQUIS-RLM LLM-as-a judge prompt (behavior-level).

```
You are evaluating an AI agent's performance
on iteration {iteration}/{max_iter}.
TASK: {query}
MEMORY STATE BEFORE: {mem_before}
THINK: {think_text}
ACT: {code}
OBSERVE: {observe}
MEMORY STATE AFTER: {mem_after}
Rate each dimension 1-5 with ONE sentence
justification.
## Core dimensions:
1. Reasoning (1-5): Did THINK show sound
reasoning based on memory?
2. Action (1-5): Was the chosen action
relevant and logical?
3. Granularity (1-5): One focused step, or
too much at once?
4. Progress (1-5): Did this iteration
meaningfully advance the task?
## Efficiency breakdown (5 sub-scores):
5a. Eff_Redundancy (1-5) — avoided repeating
a tool call?
5b. Eff_Think_Conciseness (1-5) — THINK tight
and non-repetitive?
5c. Eff_Code_Minimality (1-5) — minimal code
for its purpose?
5d. Eff_Output_Waste (1-5) — avoided
producing useless output?
5e. Eff_Tool_Choice (1-5) — most
cost-effective tool for this sub-goal?
```

```
Format EXACTLY: one line per dimension as
'Name: <score> — <reason>', then 'TOTAL:
<sum>/45'.
```

TRACE: Evidence Grounding-Guided Multi-Video Event Understanding and Claim Generation

Pengyu Yan[1,*], Akhil Gorugantu[1,*], Mahesh Bhosale[1], Abdul Wasi[1], Vishvesh Trivedi[2], David Doermann[1]

[1]University at Buffalo, SUNY, [2]New York University

Correspondence: pyan4@buffalo.edu

Abstract

Multi-video event understanding demands models that can locate and attribute query-relevant evidence scattered across long, heterogeneous video corpora. Existing large vision–language models (LVLMs) often underperform in this regime because they quickly exhaust their context budget and struggle to precisely localize evidentially important segments, frequently missing dense informational cues such as broadcast graphics, subtitles, and scoreboards. We introduce TRACE, an evidence grounding-guided framework that follows a *ground-before-reasoning* strategy for multi-video event reasoning. Our approach first builds a structured, text-searchable timeline for each video using OCR and object detection. A text-only LLM then conducts query-aware evidence localization, selecting relevant moments prior to any downstream visual reasoning. The retrieved frames and their grounding summaries are subsequently used to steer LVLM-based claim generation and cross-video citation consolidation. Experiments on MAGMaR 2026 and WikiVideo demonstrate that structured grounding markedly boosts factual completeness and attribution fidelity. On the MAGMaR validation split, TRACE raises macro-average MiRAGE F1 from 0.705 to 0.811 compared to an unguided Qwen3-VL-30B baseline, with especially strong improvements in citation recall (0.440 → 0.628). The method also attains state-of-the-art results on the official MAGMaR 2026 leaderboard.

1 Introduction

Multi-video event understanding requires models not only to recognize visual content, but to identify and attribute the specific pieces of evidence that answer a user's information need. Unlike conventional video captioning, event-centric queries often depend on sparse yet highly informative moments distributed across long collections of heterogeneous videos: a casualty count appearing briefly in a news ticker, a vote total displayed on a broadcast overlay, or an evacuation statistic mentioned alongside supporting footage. Generating factual, grounded claims from such collections is therefore fundamentally an evidence localization problem before it is a generation problem.

Recent large vision–language models (LVLMs) have demonstrated strong capabilities in generic video understanding, yet they remain poorly suited for this setting. When prompted directly with raw video, LVLMs tend to allocate attention toward visually salient content rather than query-relevant evidence, producing broad narrative summaries instead of precise, attributable claims (Martin et al., 2025a). At the same time, long-video understanding remains constrained by context capacity: even modern LVLMs can process only a limited number of frames, forcing aggressive temporal subsampling that frequently omits the brief moments containing critical information (Wu et al., 2024; Song et al., 2024). Scaling context windows alone does not resolve this bottleneck, because the challenge is not merely seeing more frames, but identifying which frames matter.

We observe that event videos contain a rich source of lightweight semantic grounding signals that existing LVLM pipelines largely underutilize. Broadcast overlays, captions, scoreboards, banners, and object co-occurrence patterns often encode the exact entities, statistics, locations, and activities required to answer factual queries. In many cases, these structured signals are more semantically informative than the raw visual appearance itself. Crucially, such signals can be extracted efficiently through OCR and object detection without invoking expensive visual reasoning (Team et al., 2025; Tian et al., 2025).

Motivated by this observation, we propose a

*Equal Contribution

Proceedings of the 2nd Workshop on Multimodal Augmented Generation via Multimodal Retrieval, pages 120–129

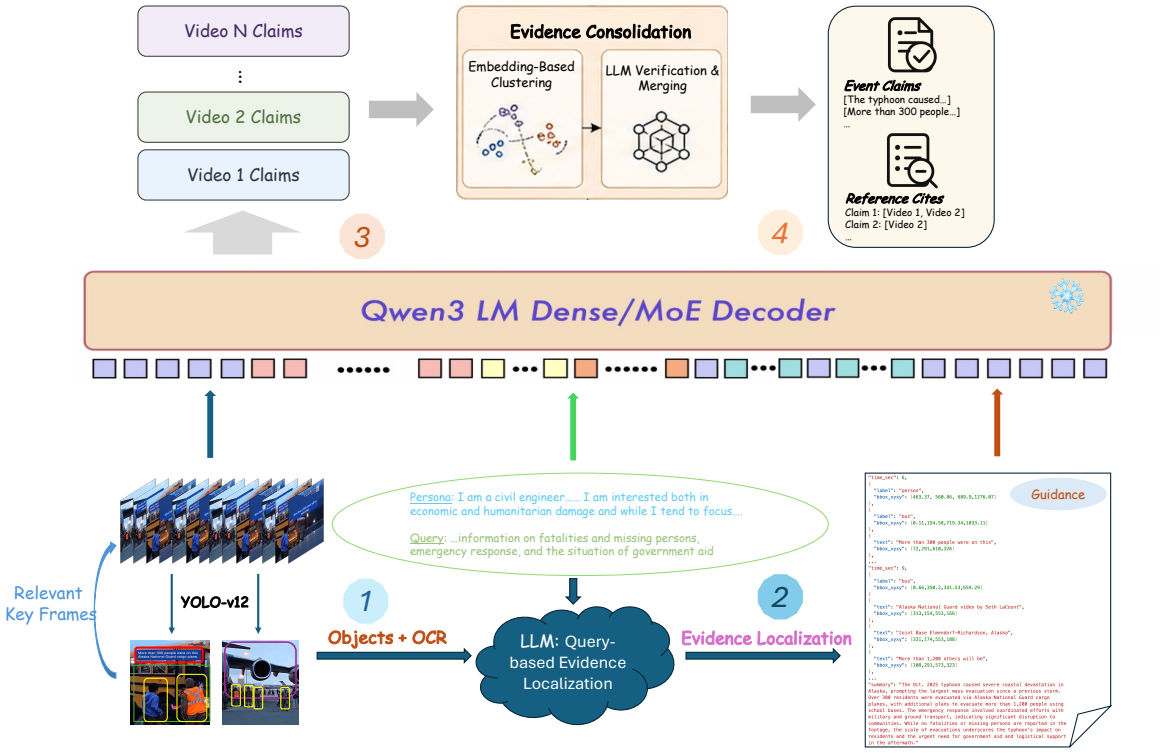


Figure 1: **Grounding-guided pipeline for event video claim generation.** We extract structured grounding signals via object detection and OCR over video frames, then use a text-only LLM to align detected labels and on-screen text with the query and persona to identify relevant moments. This text-based grounding bridges the gap between coarse detector outputs and precise query intent, producing structured guidance that directs the LVLM to relevant timestamps and conditions claim generation on explicit evidence, resulting in factual, well-grounded claims with video citations.

grounding-before-reasoning paradigm for multi-video event understanding. Instead of asking an LVLM to jointly discover evidence and generate claims from raw video, we first construct a structured, text-searchable representation of each video using OCR and object detection. A text-only LLM then aligns this grounding timeline with the query and persona to identify evidentially relevant moments and synthesize semantic guidance before any visual generation occurs. The downstream LVLM subsequently operates on a targeted subset of frames conditioned on explicit grounding cues, while a final aggregation stage consolidates claims and citations across videos.

This design directly addresses the two central failure modes of current LVLM systems. Query-conditioned grounding concentrates visual capacity on evidentially relevant moments, mitigating context saturation, while structured OCR and detection cues redirect attention toward semantically meaningful content instead of dominant visual patterns. Because the grounding stage is lightweight, text-serializable, and interpretable, it also provides a scalable alternative to brute-force long-context video processing.

We evaluate our approach on the MAGMaR 2026 Oracle Track and WikiVideo benchmarks (Martin et al., 2025a). Our method achieves state-of-the-art performance on the MAGMaR leaderboard, improving macro-average MiRAGE F1 by 8.2% over the strongest unguided Qwen3-VL baseline, with especially large gains in citation recall. The same pipeline also generalizes effectively to WikiVideo, demonstrating that lightweight structured grounding transfers across datasets and event domains.

Our contributions are summarized as follows:

- We introduce a grounding-guided pipeline that constructs a structured, text-searchable video timeline through OCR and object detection, enabling query-conditioned evidence localization prior to expensive visual reasoning.
- We propose a ground-before-reasoning paradigm that separates evidence discovery from multimodal generation, improving both

factual completeness and citation attribution in multi-video event understanding.

- We design a hybrid grounding and aggregation framework that combines targeted keyframe selection with cross-video claim deduplication and citation propagation.
- We achieve state-of-the-art results on the MAGMaR 2026 benchmark and demonstrate strong generalization on WikiVideo, with particularly large improvements in citation recall.

2 Related Work

2.1 Multi-Video Event Understanding

Recent benchmarks have shifted video understanding from generic captioning and QA toward event-centric reasoning over large collections of heterogeneous videos. MultiVENT (Sanders et al., 2023; Kriz et al., 2025) introduces multilingual event retrieval across diverse broadcast and user-generated sources, while WikiVideo (Martin et al., 2025a) studies grounded article generation from multiple event videos. These benchmarks highlight a core challenge of multi-video understanding: relevant evidence is often temporally sparse and distributed across many partially redundant sources. Our work focuses on this evidence localization bottleneck and proposes a grounding-guided framework for query-conditioned claim generation and attribution.

2.2 Long-Context Video Understanding

Large vision–language models (LVLMs) such as Video-LLaVA (Lin et al., 2024), VideoChat (Li et al., 2024), and Qwen3-VL (Team, 2025a) have become the dominant paradigm for video understanding. However, long-video reasoning remains fundamentally constrained by limited visual context capacity. Existing works address this challenge through memory compression (Song et al., 2024), adaptive frame selection (Tang et al., 2025), hierarchical representations (Ma et al., 2024), and sparse temporal sampling (Wu et al., 2024). Recent benchmarks including Video-MME (Fu et al., 2024) further demonstrate that uniformly sampled frames frequently miss short but information-dense moments critical for downstream reasoning. Our work similarly targets long-context reasoning, but approaches the problem from an evidence-grounding perspective rather than purely improving visual memory or temporal scaling.

2.3 Query-Guided Localization and Multimodal Grounding

A large body of work studies grounding language queries to temporally localized video evidence. Temporal localization and moment retrieval approaches such as Moment-DETR (Lei et al., 2021a), QVHighlights (Lei et al., 2021b), and referential video understanding systems (Qiu et al., 2024) aim to identify video segments relevant to natural-language queries. In parallel, multimodal grounding approaches including GroundingDINO (Liu et al., 2023), GLIP (Li et al., 2022), and Kosmos-2 (Peng et al., 2023) align textual semantics with visual entities and regions. Our work differs from these approaches in that we use lightweight OCR and object detections as structured semantic grounding signals for downstream evidence routing and claim generation across multiple videos.

2.4 OCR and Structured Semantic Signals

Event videos contain rich structured semantic cues embedded in overlays, captions, scoreboards, and broadcast graphics. OCR-based multimodal reasoning benchmarks such as TextVQA (Singh et al., 2019), ST-VQA (Biten et al., 2019), ChartReformer (Yan et al., 2024), and OCR-VQA (Mishra et al., 2019) demonstrate the importance of scene text for factual visual understanding. Meanwhile, OCR systems including PaddleOCR (Du et al., 2020) and HunyuanOCR (Tencent Hunyuan Team, 2025) provide efficient extraction of textual evidence from visual content. Similar interactions between graphical structure and embedded text have also been explored in document and chart understanding (Yan et al., 2024). Our framework extends these ideas to long-video event understanding, where OCR often provides more semantically precise evidence than raw visual appearance alone.

2.5 Retrieval-Augmented and Modular Multimodal Reasoning

Recent multimodal systems increasingly separate evidence discovery from downstream reasoning through retrieval-augmented or modular architectures. Retrieval-augmented language models (Lewis et al., 2020; Borgeaud et al., 2022; Asai et al., 2024) improve factuality by retrieving supporting evidence prior to generation, while modular multimodal systems such as Visual Programming (Gupta and Kembhavi, 2023),

ViperGPT (Suris et al., 2023), HuggingGPT (Shen et al., 2023), and MM-REACT (Yang et al., 2023) demonstrate the effectiveness of decomposing perception and reasoning into specialized stages. Our work extends this paradigm to multi-video event understanding by introducing a grounding-guided framework that performs lightweight semantic evidence localization before expensive multimodal reasoning.

3 Method

Our goal is to generate factual, query-conditioned claims from a collection of event videos while preserving explicit attribution to supporting sources. Rather than relying on an LVLM to jointly discover evidence and perform generation directly from raw video, we decompose the task into two main stages: lightweight evidence grounding followed by grounding-guided multimodal reasoning.

The central idea of our approach is to transform long, unstructured videos into a structured semantic representation that can be efficiently searched and filtered prior to expensive visual inference. We first extract lightweight grounding signals through OCR and object detection to construct a text-searchable timeline of each video. A text-only LLM then performs query-conditioned evidence localization over this timeline, identifying the moments most relevant to the user query and persona. Finally, an LVLM generates claims conditioned on both the selected frames and their associated semantic guidance. Claims from multiple videos are subsequently consolidated through cross-video evidence aggregation. An overview of the pipeline is shown in Figure 1.

3.1 Structured Video Grounding

Long event videos contain large amounts of redundant visual content interspersed with sparse but highly informative evidence-bearing moments. Processing all frames uniformly with an LVLM is both computationally inefficient and poorly aligned with the evidence localization nature of the task. We therefore first convert each video into a lightweight structured grounding representation that can be queried efficiently before downstream visual reasoning.

Object detection. YOLOv12 (Tian et al., 2025) processes each sampled frame, yielding per-frame detections $\mathcal{D}_t = \{(l_i, c_i, \mathbf{b}_i)\}_i$, where $l_i \in \mathcal{L}_{\text{COCO-80}}$, $c_i \in [0, 1]$, and $\mathbf{b}_i$ is the axis-aligned bounding box. Object co-occurrence patterns carry rich contextual signal beyond individual labels: the simultaneous presence of `person`, `microphone`, and `podium` reliably identifies a press-conference segment without any scene-level supervision.

Text recognition. An OCR module extracts visible text strings from each frame. Broadcast lower-thirds, scoreboards, and graphical overlays name entities, statistics, and locations that object detectors cannot recover, making on-screen text the highest-precision signal in news and event footage.

The two streams are merged into a chronological timeline

$$\mathcal{F} = \left\{(t,\, \mathcal{D}_t,\, \mathcal{T}_t)\right\}_{t=0}^{T}. \tag{1}$$

Because $\mathcal{F}$ is fully text-serializable, the subsequent grounding step is vision-free — fast, deterministic, and interpretable.

3.2 Query-Conditioned Evidence Localization

The structured grounding timeline provides dense semantic coverage of the video, but only a small subset of frames are typically relevant to a given query and persona. Rather than performing expensive multimodal reasoning over the entire video, we first localize evidentially relevant moments using a lightweight text-only reasoning stage.

Evidence localization. A key challenge is that low-level detector outputs do not naturally align with open-ended user queries. Relevant evidence is often expressed indirectly through combinations of OCR text, object co-occurrence, and contextual cues rather than explicit keyword matches. For example, an election-related query may correspond to frames containing vote percentages, podium scenes, and broadcast overlays even when no detector label directly references elections. We therefore introduce a query-conditioned grounding stage that bridges the gap between perception outputs and semantic intent. The timeline $\mathcal{F}$ is partitioned into non-overlapping windows $\{\mathcal{F}_j\}$ of C consecutive frames. Each window is serialized into a compact textual representation containing timestamps, detected objects, and OCR text. And they are prompted to the LLM alongside q and p; the model returns the relevant subset $\mathcal{S}_j \subseteq \mathcal{F}_j$ together with the supporting detections and OCR strings. The union

$$\mathcal{S} = \bigcup_j \mathcal{S}_j \tag{2}$$

constitutes the query-relevant keyframe set for that video. Importantly, this stage operates entirely in text space without invoking a vision encoder, making evidence localization substantially more efficient than dense LVLM inference.

Grounding summary. Frame-level detections and OCR signals provide sparse semantic anchors, but downstream LVLM generation still requires higher-level contextual understanding of how these observations relate to the query and persona. We therefore introduce an intermediate grounding-summary stage that compresses localized evidence into a coherent semantic description prior to visual generation. This summary acts as a semantic bridge between low-level perception outputs and downstream multimodal reasoning, transforming fragmented detector observations into an interpretable representation of the underlying event narrative.

3.3 Grounding-Guided Claim Generation

After evidence localization, the downstream LVLM performs claim generation conditioned on both the original video content and the structured grounding signals.

Hybrid frame selection. We construct the LVLM input using a hybrid frame-selection strategy that combines uniformly sampled frames with guidance-targeted evidence frames.

The visual input to the LVLM is the union

$$\mathcal{I}_v = \mathcal{I}_{\text{unif}} \cup \{\hat{i}_s : t_s \in \mathcal{S}\}, \tag{3}$$

where $\mathcal{I}_{\text{unif}}$ comprises N_{unif} linearly spaced frames for broad narrative coverage, and each relevant timestamp t_s is mapped to its nearest frame index

$$\hat{i}_s = \min\Big(\lfloor t_s \cdot \text{fps} \rceil,\ F_{\text{total}} - 1\Big). \tag{4}$$

After deduplication, frames are sorted temporally and decoded at 448×448 pixels. The uniform sampling preserves broad temporal coverage and guards against potential errors and noise introduced during grounding, while targeted frames allocate visual capacity toward moments identified as evidentially relevant.

Temporal alignment. Frame indices (Eq. 4) are passed as explicit positional metadata rather than dense ranks $0, 1, \ldots, N-1$. This preserves correct temporal spacing in the model's rotary position embeddings, letting the LVLM correlate textual grounding annotations (e.g., "t=45 s: on-screen seat count") with their visual tokens. Without this alignment, the text and visual temporal axes diverge, undermining cross-modal grounding.

Evidence fusion. The five evidence streams are assembled into a single prompt and passed to the LVLM to generate per-video claims:

$$\mathcal{C}_v = \text{LVLM}(\mathcal{I}_v,\ q,\ p,\ \mathcal{A}_{\mathcal{S}},\ g,\ \text{ASR}_v)\,, \tag{5}$$

where $\mathcal{A}_{\mathcal{S}}$ denotes the structured frame-level annotations derived from $\mathcal{S}$ — each entry recording the timestamp, detected objects, and OCR strings of a relevant keyframe. The remaining inputs are the hybrid frame set $\mathcal{I}_v$, the query q and persona p, the grounding summary g, and the Whisper ASR transcript ASR_v. Annotations are cast as *supplementary grounding hints* that the model must cross-validate against the video, preventing over-reliance on potentially noisy detector outputs. The model is instructed to output single-sentence claims grounded in directly observed evidence, with a preference for specific facts (names, numbers, dates) over vague paraphrases.

3.4 Cross-Video Claim Consolidation

We frame aggregation as a cross-video evidence consolidation problem rather than a simple textual deduplication task. The goal is not merely to suppress repeated claims, but to reconcile semantically equivalent evidence across videos while preserving the full set of supporting sources.

To achieve this, we first encode generated claims into a semantic embedding space and perform conservative similarity-based clustering. Candidate clusters are subsequently verified by an LLM operating under a strict same-proposition criterion, allowing the system to distinguish genuine paraphrases from superficially similar but factually distinct claims. For each cluster, we retain the most information-complete claim as the canonical representation and propagate the union of supporting video citations across all cluster members. This strategy improves citation recall by explicitly consolidating evidence distributed across multiple videos while avoiding the precision degradation associated with aggressive generative merging.

4 Experiments

4.1 Implementation Details

Models and hardware. All pipeline stages run on four NVIDIA RTX A6000 GPUs (48 GB each;

Table 1: **Official MAGMaR 2026 Leaderboard** (best submission per team, selected entries). We calculate the F1 and Avg. F1 based on the Info/Cite P/R offered by the MAGMaR 2026 workshop. Our results leads all teams on all Recall and F1 measures and ranks second in human evaluation, trailing the top team by only 0.008 points. The baseline model is CAG in (Martin et al., 2025a)

Team	Human Evaluation	Best Votes	Avg. F1	Reference Info			Reference Cite		
				P	R	F1	P	R	F1
HAIVLab	2.526	2	0.455	0.584	0.450	0.508	0.479	0.347	0.402
CiteChasers	2.542	0	0.349	0.609	0.304	0.406	0.509	0.204	0.291
MARS-Bullet	2.667	0	0.424	0.711	0.394	0.507	0.604	0.237	0.340
MARS-ss-qa-base	3.070	6	0.299	0.331	0.306	0.318	0.277	0.281	0.279
Baseline (CAG)	3.088	1	0.434	0.764	0.410	0.534	0.617	0.228	0.333
MARS-Ginger	3.123	6	0.433	**0.776**	0.404	0.531	**0.643**	0.226	0.334
MARS-RLM	3.298	3	0.436	0.708	0.385	0.499	0.592	0.272	0.373
MARS-iter-qa-ginger	3.694	5	0.278	0.345	0.290	0.315	0.257	0.226	0.241
MARS-ss-qa-ginger	3.421	**10**	0.341	0.544	0.324	0.406	0.326	0.238	0.275
MARS-iter-qa-base	**3.833**	8	0.296	0.347	0.313	0.329	0.268	0.258	0.263
Ours	3.825	8	**0.499**	0.640	**0.483**	**0.551**	0.498	**0.405**	**0.447**

192 GB total VRAM). The LLM stages — temporal grounding filter and cross-video aggregation — use **Qwen3-30B-A3B-Instruct** (Team, 2025b) in BF16 precision, while the LVLM claim generation stage uses **Qwen3-VL-30B-A3B-Instruct** (Team, 2025b) in BF16. Both models are served via vLLM with tensor parallelism across all four GPUs and are loaded sequentially, so the full 192 GB budget is available to each stage.

Token budget. Frames are resized to 448×448 pixels, yielding approximately 256 visual tokens per frame under Qwen3-VL's visual tokenizer. With $N_{\text{unif}} = 100$ uniform frames and at most 30 guidance-targeted keyframes, the visual token ceiling is $130 \times 256 = 33{,}280$. Text context — query, persona, frame annotations, and ASR transcript — contributes approximately 3,600 additional tokens, placing a typical prompt at $\sim$29,000 tokens, comfortably within the 32,768-token context window.

4.2 Experimental Setup

Datasets. Our primary benchmark is the **MAGMaR 2026 Oracle Track** validation set, comprising 8 event topics drawn from real-world news events. Each topic is paired with a curated set of relevant videos and gold claims annotated with per-claim video citations. To assess generalization, we additionally evaluate on the **WikiVideo** dataset, which contains 52 queries paired with multi-video collections spanning diverse topics, which is 398 unique videos in total, using the same pipeline and evaluation protocol.

Evaluation metrics. For automatic evaluation, MiRAGE (Martin et al., 2025b) assesses predictions along two axes: **Reference Info** (InfoP/R), measuring factual completeness of predicted claims against the gold set, and **Reference Cite** (CiteP/R), measuring accuracy of per-claim video citations. Each entailment judgment within MiRAGE is produced by CLUE (Zhang et al., 2026). We compute F1 scores from the reported precision and recall via the harmonic mean, and additionally report **Avg. F1**, the macro-average of InfoF1 and CiteF1, as a single summary statistic. In MAGMaR workshop, human evaluation is conducted from three annotators scoring each system on a 1–5 scale across five dimensions: factuality, adequacy, coherence, relevancy, and fluency; they additionally select the single best response per query as vote number.

4.3 Official Workshop Results

Table 1 presents the official MAGMaR 2026 workshop leaderboard. Our results achieves the highest scores on most automatic metrics, and achieves the highest final F1 score: InfoF1 0.551, CiteF1 0.447, and Avg. F1 0.499, exceeding the second-ranked team (HAIVLab) by **+0.049** in **Avg. F1**. Notably, our Avg. F1 also surpasses the workshop-provided CAG baseline (Martin et al., 2025a) by **+0.065**, which achieves the second-highest InfoP (0.764) among all teams despite its lower recall.

In human evaluation — where annotators score factuality, adequacy, coherence, relevancy, and fluency on a 1–5 scale — we rank second with 3.825, trailing MARS-iter-qa-base by only 0.008 while matching their tally of 8 "best" votes. Notably, MARS-ss-qa-ginger receives the most best votes (10) despite ranking lower in both scalar human score and automatic metrics, suggesting that pair-

Table 2: **Comparison with LVLM baselines** on the MAGMaR 2026 Oracle Track validation set (8 topics). Our grounding-guided system achieves the highest Avg. F1 (0.811), with gains concentrated in citation recall.

Method	Avg. F1	Reference Info			Reference Cite		
		P	R	F1	P	R	F1
Qwen3.5-9B	0.472	0.437	0.756	0.554	0.875	0.251	0.390
Qwen3-VL-8B	0.723	0.870	0.802	0.835	0.93	0.452	0.608
Qwen3-VL-30B	0.705	**0.883**	0.731	0.800	**0.990**	0.440	0.609
Ours	**0.811**	0.863	**0.876**	**0.869**	0.939	**0.628**	**0.753**

Table 3: **Generalization to WikiVideo** (52 queries, 398 videos). Avg. F1 is the macro-average of InfoF1 and CiteF1. Our pipeline maintains the highest Avg. F1 and citation recall, consistent with MAGMaR findings.

Metric	Qwen3-VL-8B	Qwen3-VL-30B	Ours
Avg. F1	0.878	0.854	**0.879**
Reference Info			
P	**0.915**	0.888	0.868
R	0.885	0.905	**0.918**
F1	**0.885**	0.880	0.882
Reference Cite			
P	0.991	**0.993**	0.936
R	0.792	0.767	**0.838**
F1	0.871	0.828	**0.876**

wise preference captures a distinct quality dimension from scalar ratings. The strong alignment between our automatic metric leadership and near-top human evaluation provides evidence that the MiRAGE framework is a reliable proxy for human judgment on this task.

4.4 Comparison with LVLM Baselines

Baselines. We compare against three LVLM baselines that receive no grounding guidance. **Qwen3.5-9B** is a compact vision–language model applied directly to the video and query–persona context. **Qwen3-VL-8B** is a medium-scale VLM baseline using uniform frame sampling only. **Qwen3-VL-30B** shares the identical backbone with our pipeline but is prompted with video frames and query–persona context alone, without object detection, OCR, or any LLM grounding filter, directly isolating the contribution of our multi-modal grounding stage.

Since ground truth annotations are unavailable for the full workshop test set, we conduct controlled comparisons on the MAGMaR validation subset (8 topics), for which gold claims and per-claim citations are provided. Table 2 reports MiRAGE scores on this set. Our grounding-guided pipeline outperforms all baselines across every metric, with the best configuration achieving Avg. F1 of **0.811** versus 0.705 for the strongest baseline (Qwen3-VL-30B), a gain of $+0.106$.

Citation recall is the primary bottleneck for unguided models. Qwen3-VL-30B achieves high citation precision (0.990) but very low recall (0.440): without grounding, the model anchors on the most salient video while overlooking the broader evidence base. Our structured guidance record raises CiteR from 0.440 to **0.628** ($+42.7\%$ relative) and CiteF1 from 0.609 to **0.753**, demonstrating that the grounding stage directs the model to cite the full range of relevant sources. Citation precision remains high at 0.939, confirming that the additional citations are well-grounded rather than spurious.

Factual completeness also improves substantially. InfoF1 rises from 0.800 (Qwen3-VL-30B) to **0.869** under our best configuration, reflecting that grounding-conditioned generation produces claims that more thoroughly cover the gold annotation. This gain is consistent across all four of our variants (≥ 0.859), confirming that it stems from the grounding stage itself rather than from any particular downstream choice.

4.5 Generalization to WikiVideo

Table 3 evaluates the same pipeline on WikiVideo, a larger and more diverse dataset with 52 queries. Our method achieves Avg. F1 of **0.879**, edging both Qwen3-VL-8B (0.878) and Qwen3-VL-30B (0.854). The pattern of gains mirrors MAGMaR: citation recall improves most ($0.792 \rightarrow \mathbf{0.838}$) and CiteF1 remains the highest among all methods ($\mathbf{0.876}$). The smaller absolute margins on WikiVideo reflect the already-high baseline performance on this dataset.

We attribute the reduced performance gap to two

Table 4: **Ablation study** on the MAGMaR 2026 Oracle Track validation set (8 topics), examining the effect of guided keyframe augmentation and aggregation strategy. Embedding-based aggregation and keyframe augmentation provide complementary gains, with their combination achieving the best overall result.

Additional Key Frames	Aggregation Method	Avg. F1	Reference Info			Reference Cite		
			P	R	F1	P	R	F1
✗	LLM	0.802	0.860	0.858	0.859	0.921	0.626	0.745
✗	Embed-Sim	0.808	0.862	0.873	0.868	0.925	**0.628**	0.748
✓	LLM	0.804	0.849	**0.885**	0.867	0.931	0.616	0.741
✓	Embed-Sim	**0.811**	0.863	0.876	**0.869**	0.939	**0.628**	**0.753**

structural differences between the datasets. First, WikiVideo videos are considerably shorter (mean 60.1 s, median 55.3 s) compared to MAGMaR (mean 104.9 s, median 58.4 s), and their duration distribution is more uniform (std 47.6 s vs. 120.8 s). For shorter, temporally compact videos, uniform frame sampling already provides dense coverage of the visual content, diminishing the marginal benefit of our YOLO- and OCR-guided keyframe selection. Second, unguided baselines already achieve near-ceiling performance on WikiVideo (Avg. F1 ≥ 0.854), leaving limited headroom for further improvement. Together, these factors explain why the grounding advantage observed on MAGMaR ($+0.106$ over Qwen3-VL-30B) does not fully transfer to WikiVideo ($+0.025$), while the consistent citation recall gain ($+0.046$ CiteR) confirms that multi-modal grounding remains beneficial even in this easier regime.

4.6 Ablation Study

Our variants. We evaluate four configurations of our pipeline varying two dimensions: (i) **Frame selection** — uniform 100 frames only (✗) versus uniform frames augmented with guidance-targeted keyframes (✓); and (ii) **Aggregation** — LLM-based cross-video merging (LLM) versus embedding-similarity deduplication with LLM verification (EMBED-SIM).

Table 4 breaks down the contribution of each pipeline component. All four variants comfortably exceed the strongest baseline (Avg. F1 ≥ 0.802 vs. 0.705), confirming that multi-modal grounding is the dominant source of improvement regardless of downstream configuration.

Embedding-based aggregation is consistently better. EMBED-SIM aggregation outperforms LLM aggregation in both frame-selection settings ($+0.006$ and $+0.007$ in Avg. F1, respectively). The advantage is most visible in CiteF1 (0.748 vs. 0.745; 0.753 vs. 0.741), suggesting that similarity-based deduplication is more precise at suppressing redundant claims than purely generative merging.

Guided keyframe augmentation provides complementary gains. Adding guidance-targeted keyframes (✓) improves InfoR from 0.858 to **0.885** under LLM aggregation, indicating that the additional frames expose query-relevant visual evidence missed by uniform sampling. Gains are modest under EMBED-SIM ($+0.003$ Avg. F1), suggesting that the text-based grounding signal already captures much of this context at the prompt level. The combination of guided keyframes and EMBED-SIM aggregation yields the best overall result: Avg. F1 **0.811**, InfoF1 **0.869**, and CiteF1 **0.753**.

5 Conclusion

We presented a grounding-guided pipeline for multi-video event claim generation that adopts a ground-then-generate paradigm: lightweight detection and OCR signals direct a text-only LLM to query-relevant keyframes before any visual inference, and a downstream LVLM generates attributed claims conditioned on the resulting guidance. The approach consistently outperforms unguided LVLM baselines on both MAGMaR 2026 and WikiVideo, with the largest gains in citation recall — confirming that structured perception-based grounding is an effective and transferable principle for video claim attribution.

Limitations and future work. The pipeline's object detector is constrained to the COCO-80 vocabulary, limiting its ability to identify domain-specific entities central to many news queries. The sequential, non-differentiable design also means grounding errors propagate without recovery. Future directions include open-vocabulary detection (Liu et al., 2023), adaptive frame sampling for fast-paced events, timestamp-level citation attribution, and end-to-end joint optimization of the grounding and generation stages.

References

Akari Asai and 1 others. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *ICLR*.

Ali Furkan Biten and 1 others. 2019. Scene text visual question answering. In *ICCV*.

Sebastian Borgeaud and 1 others. 2022. Improving language models by retrieving from trillions of tokens. *ICML*.

Yuning Du and 1 others. 2020. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941*.

Chaoyou Fu and 1 others. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *CVPR*.

Reno Kriz, Kate Sanders, David Etter, Kenton Murray, Cameron Carpenter, Carl Van Ess-Dykema, Eugene Yang, Hamed Sayyed, Robin Carmody, Mahsa Roberts, and Benjamin Van Durme. 2025. MultiVENT 2.0: A massive multilingual benchmark for event-centric video retrieval. *arXiv preprint arXiv:2410.11619*. Verify exact arXiv ID and author list on Scholar.

Jie Lei and 1 others. 2021a. Moment-detr: End-to-end video moment retrieval and highlight detection. In *NeurIPS*.

Jie Lei and 1 others. 2021b. Qvhighlights: Detecting moments and highlights in videos via natural language queries. In *NeurIPS*.

Patrick Lewis and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2024. VideoChat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Liunian Harold Li and 1 others. 2022. Glip: Grounded language-image pre-training. In *CVPR*.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of EMNLP*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2023. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Fanqing Ma and 1 others. 2024. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*.

Alexander Martin, Kate Sanders, William Walden, Dengjia Zhang, Reno Kriz, Angela Cao, Adarsh Pyarelal, Eugene Yang, and Benjamin Van Durme. 2025a. WikiVideo: Article generation from multiple videos. *arXiv preprint arXiv:2504.00939*.

Alexander Martin, William Walden, Reno Kriz, Dengjia Zhang, Kate Sanders, Eugene Yang, Chihsheng Jin, and Benjamin Van Durme. 2025b. Seeing through the mirage: Evaluating multimodal retrieval augmented generation. *Preprint*, arXiv:2510.24870.

Anand Mishra and 1 others. 2019. Ocr-vqa: Visual question answering by reading text in images. *ICDAR*.

Nanyun Peng and 1 others. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Jihao Qiu, Yuan Zhang, Xi Tang, Lingxi Xie, Tianren Ma, Pengyu Yan, David Doermann, Qixiang Ye, and Yunjie Tian. 2024. Artemis: Towards referential understanding in complex videos. *arXiv preprint arXiv:2406.00258*.

Kate Sanders, David Etter, Reno Kriz, and Benjamin Van Durme. 2023. Multivent: Multilingual videos of events and aligned natural text. *Advances in Neural Information Processing Systems*, 36:51065–51079.

Yongliang Shen and 1 others. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *NeurIPS*.

Amanpreet Singh and 1 others. 2019. Towards vqa models that read. In *CVPR*.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. 2024. MovieChat: From dense token to sparse memory for long video understanding. In *Proceedings of CVPR*.

Dídac Suris and 1 others. 2023. Vipergpt: Visual inference via python execution for reasoning. *ICCV*.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. 2025. Adaptive keyframe sampling for long video understanding. *arXiv preprint arXiv:2502.21271*.

Hunyuan Vision Team, Pengyuan Lyu, Xingyu Wan, Gengluo Li, Shangpin Peng, Weinong Wang, Liang Wu, Huawen Shen, Yu Zhou, Canhui Tang, Qi Yang, Qiming Peng, Bin Luo, Hower Yang, Xinsong Zhang, Jinnian Zhang, Houwen Peng, Hongming Yang, Senhao Xie, and 12 others. 2025. Hunyuanocr technical report.

Qwen Team. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Qwen Team. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Tencent Hunyuan Team. 2025. HunyuanOCR: A multilingual OCR model from tencent Hunyuan. Tencent Hunyuan Team. Model card available at https://huggingface.co/tencent/HunyuanOCR; verify final citation form.

Yunjie Tian, Qixiang Ye, and David Doermann. 2025. YOLOv12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*.

Pengyu Yan and 1 others. 2024. Chartreformer: Natural language-driven chart image editing. *arXiv preprint arXiv:2403.00209*.

Zhengyuan Yang and 1 others. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Dengjia Zhang, Alexander Martin, William Jurayj, Kenton Murray, Benjamin Van Durme, and Reno Kriz. 2026. Unified multimodal uncertain inference. *Preprint*, arXiv:2604.08701.

CRAFT: Critic-Refined Adaptive Key-Frame Targeting for Multimodal Video Question Answering

Mahesh Bhosale[1*†] Abdul Wasi[1*] Vishvesh Trivedi[2*]
Pengyu Yan[1] Akhil Gorugantu[1] David Doermann[1]

[1]University at Buffalo [2]New York University

Abstract

Grounded multi-video question answering over real-world news events requires systems to surface query-relevant evidence across heterogeneous video archives while attributing every claim to its supporting source. We introduce CRAFT (Critic-Refined Adaptive Key-Frame Targeting), a query-conditioned pipeline that combines dynamic keyframe selection, per-video ASR with multilingual fallback, and a hybrid critic loop to iteratively verify and repair claims before consolidation. The pipeline integrates UNLI temporal entailment, DeBERTa-v3 cross-claim screening, and a Llama-3.2-3B adjudicator, with a final citation-merging stage that emits each fact once with all supporting source identifiers. On MAGMaR 2026, CRAFT achieves the best overall average (0.739), reference recall (0.810), and citation F1 (0.635). We further evaluate on a MAGMaR-style conversion of WikiVideo with 52 non-overlapping event queries, where CRAFT also performs strongly (0.823 Avg), showing that its claim-centric evidence aggregation generalizes beyond MAGMaR. Ablations show that atomic claims, ASR, and the critic loop drive the main gains over the vanilla query-conditioned baseline. Code and implementation details are publicly available at https://github.com/bhosalems/CRAFT.

1 Introduction

Multi-video question answering over real-world news events underlies tasks from event understanding to fact-checking and crisis reporting. Recent benchmarks such as MultiVENT 2.0 (Kriz et al., 2025) and the WikiVideo article-generation task (Martin et al., 2025a) formalize a strict variant of this problem: given a query and a collection of relevant videos, a system must produce a report whose every statement is grounded in identifiable visual, textual, or spoken evidence from the source videos. The MAGMaR 2026 oracle task adds two further constraints. Each query is paired with a persona and a background paragraph, and the resulting report is scored on six axes that separately measure content precision and recall (REF-P, REF-R) against a reference answer and citation precision and recall (CITE-P, CITE-R) against gold source videos. A high-scoring system must both surface the right facts and attach the right videos to them.

Three properties of long news video make this hard. First, vision-language models face a hard token-budget bottleneck on hour-scale input: even at 1 FPS, a long video exceeds practical context windows (Tang et al., 2025; Gao et al., 2025), and uniform sampling silently truncates whatever falls outside the budget. Second, even when relevant frames are presented, recent hallucination benchmarks (Wang et al., 2024b; Li et al., 2025; Zhang et al., 2024b) show that VLMs routinely emit claims unsupported by the visual content, with errors concentrated at long-tail entities, numerical details, and event timing-precisely the content most likely to be cited in a news report. Third, much of the answer-relevant content in news video is spoken rather than shown: visual-only extraction misses interview answers, on-the-ground reporting, and official statements, especially in non-English coverage.

Prior work addresses these challenges in isolation. Adaptive keyframe selectors (Tang et al., 2025; Gao et al., 2025) trim the visual input to query-relevant frames but treat the result as terminal evidence, with no check that downstream claims are actually supported. Critic-driven video QA systems (Liu et al., 2026; Dang et al., 2025) add verification, but typically at the final answer-aggregation stage and at the granularity of a single role rather than per claim. Modular video-RAG pipelines (Jeong et al., 2025; Ren et al., 2025; Zeng et al., 2025) compose retrieval and reasoning over

*Equal contribution.
† Correspondence: mbhosale@buffalo.edu.

Proceedings of the 2nd Workshop on Multimodal Augmented Generation via Multimodal Retrieval, pages 130–143

long context but rely on a single visual stream and ignore speech leading to citation faithfulness diverging from citation correctness (Wallat et al., 2025).

We present **CRAFT** (Critic-Refined Adaptive Key-Frame Targeting), a query-conditioned pipeline that integrates these threads for the MAGMaR 2026 oracle task (Figure 1). Our contributions are: (i) a *multimodal evidence stream* (§3.1) combining 120-second video chunking, per-video ASR (Qwen3-ASR-1.7B with a Whisper-large-v3 fallback for low-resource languages), automatic English translation, and dynamic query-conditioned keyframe selection, so the VLM receives a clip and transcript both targeted at the current query; (ii) a *critic-guided extraction loop* (§3.3) that runs a UNLI video-claim entailment model for temporal grounding, a DeBERTa-v3 MNLI cross-encoder for cross-claim contradiction screening, and a Llama-3.2-3B adjudicator that confirms contradictions and emits repair feedback, returning the critic report to the VLM for up to four re-extraction rounds; and (iii) atomic claim formatting (§3.2) with *citation-merging* consolidation (§3.6), which emits each fact once with all supporting source identifiers attached, preserving citation recall while suppressing the redundancy that inflates reference-precision loss.

On MAGMaR 2026 (§4), CRAFT outperforms strong baselines with the highest overall average (0.739), reference recall (0.810), and citation F1 (0.635) of all evaluated configurations. Ablations (§4.5) show that the gains from the critic loop, atomic claims, and ASR-augmented extraction are partly orthogonal to the choice of base VLM, transferring across Qwen3.5-9B (Qwen Team, 2026) and Qwen3-VL-30B (Bai et al., 2025a) backbones, and outperforming strong VLMs such as Molmo2-8B (Deitke et al., 2025) and Gemma-4-31B[1].

2 Related Work

Long-video understanding with vision-language models. Open-source video-language models have improved rapidly along two axes: backbone capacity and temporal modeling. The Qwen-VL family progressed from dynamic-resolution and time-aligned M-RoPE in Qwen2.5-VL (Bai et al., 2025b) to interleaved M-RoPE, DeepStack cross-layer fusion, and explicit timestamp tokens in Qwen3-VL (Bai et al., 2025a), while InternVL3 (Zhu et al., 2025) introduced Variable Visual Position Encoding and native multimodal pre-training. LLaVA-Video (Zhang et al., 2024c) and LLaVA-OneVision (Li et al., 2024) consolidated the LLaVA recipe for video instruction tuning. Despite these gains, all such models face a hard token-budget bottleneck on hour-scale input: even at 1 FPS, a long video produces token counts that exceed practical context windows (Tang et al., 2025; Gao et al., 2025). Specialized long-context architectures, including LongVU (Shen et al., 2024), Video-XL (Shu et al., 2025), MovieChat (Song et al., 2024), and MA-LMM (He et al., 2024), mitigate this through spatiotemporal compression, sparse memory, or hierarchical attention, but typically at the cost of fine-grained temporal evidence that is essential for citation-grounded answering.

Adaptive keyframe selection. Because uniform sampling is the dominant performance bottleneck on long videos, a substantial body of recent work has focused on query-conditioned frame selection. AKS (Tang et al., 2025) formulates selection as a joint optimization over prompt-frame relevance and temporal coverage, solved by a recursive split-and-judge algorithm; APVR (Gao et al., 2025) extends this idea to a two-granularity hierarchy in which Pivot Frame Retrieval expands the query into semantic facets and Pivot Token Retrieval performs query-aware token selection within retained frames. VideoTree (Wang et al., 2025) replaces flat selection with a query-adaptive tree of clustered keyframes captioned coarse-to-fine. Other recent variants include MDP3 (Sun et al., 2025b), which casts selection as a Markov decision process; Q-Frame (Zhang et al., 2025a), which ranks frames into multiple resolution tiers; AdaRD-Key (Zhang et al., 2025b), which encourages diversity through determinantal point processes; F2C (Sun et al., 2025a), which extends keyframes to short clips to preserve motion continuity; and VidF4 (Liang et al., 2024), which proposes differentiable frame scoring for end-to-end VideoQA. A.I.R. (Zou et al., 2025) and T* (Ye et al., 2025) replace lightweight CLIP-based scorers with iterative VLM-based reasoning over candidate frames, trading cost for accuracy. A common property of these selectors is that their output is treated as the terminal evidence representation, with no mechanism to detect whether claims subsequently extracted from the chosen frames are actually supported by the video.

[1] https://huggingface.co/google/gemma-4-31B-it

Modular and agentic video pipelines. Modular pipelines decompose video question answering into captioning, retrieval, and reasoning stages. LLoVi (Zhang et al., 2024a) demonstrated that short-clip captions plus an LLM aggregator can match dedicated video models on long-form benchmarks. VideoAgent (Wang et al., 2024a) introduced an iterative agent that uses CLIP-based frame retrieval and self-reflective stopping, achieving strong results on EgoSchema and NExT-QA with fewer than ten frames on average. MoReVQA (Min et al., 2024) showed that a multi-stage event-parser, grounding, and reasoning architecture with shared external memory outperforms single-stage program-generation approaches. More recent agentic systems, including VideoAgent2 (Zhi et al., 2025), Deep Video Discovery (Zhang et al., 2025c), and VideoDeepResearch (Yuan et al., 2025), equip a reasoning model with multi-granular search tools over a structured video index. These systems generally place verification, when present at all, at the final answer-aggregation stage rather than during evidence extraction.

Critic-driven refinement and faithfulness. Several lines of work have explored verification and critic loops to improve grounding. In text generation, Self-RAG (Asai et al., 2024) and CRAG (Yan et al., 2024) introduce reflection tokens or evaluators that trigger retrieval correction. For video, VideoMind (Liu et al., 2026) defines four explicit roles-planner, grounder, verifier, and answerer-instantiated as Chain-of-LoRA adapters, and demonstrates that the verifier role substantially improves grounding accuracy. MUPA (Dang et al., 2025) runs three reasoning paths in parallel and consolidates them through a reflection agent. Wallat et al. (2025) further show that, in retrieval-augmented generation, citation correctness diverges sharply from citation faithfulness, motivating verification as a first-class component. Hallucination benchmarks for video, including VideoHallucer (Wang et al., 2024b), EventHallusion (Zhang et al., 2024b), and VidHalluc (Li et al., 2025), document that vision-language models routinely emit unsupported claims even when relevant frames are available. CRAFT builds on this line of work by applying a hybrid critic with iterative repair feedback at the claim level, at finer granularity than the single verifier role of Liu et al. (2026).

Multi-video corpora and grounded generation. At the corpus level, MultiVENT 2.0 (Kriz et al., 2025) provides a large-scale multilingual benchmark of event-centric news videos, accompanied by retrieval baselines such as MMMORRF (Samuel et al., 2025) that fuse modality-specific scores via weighted reciprocal rank fusion. WikiVideo (Martin et al., 2025a) formalizes the task of generating articles whose every claim is grounded in audio, video, or on-screen text from a video collection. VideoRAG variants (Jeong et al., 2025; Ren et al., 2025) extend retrieval-augmented generation to long-context video, while SceneRAG (Zeng et al., 2025) substitutes scene-level segmentation for fixed chunking. Our pipeline follows the claim-centric formulation introduced by WikiVideo and instantiates it for the multi-video setting with explicit citation merging at consolidation.

3 Method

We propose a query-conditioned multimodal video question answering pipeline for the MAGMaR 2026 oracle task, where each query is paired with a set of relevant videos. This setting differs from standard single-video VQA because the answer may require evidence distributed across multiple videos. Moreover, irrelevant or redundant clips can easily introduce unsupported claims. Our pipeline, similar to Martin et al. (2025a), therefore follows a claim-centric design: it first extracts atomic, source-grounded claims from each query-video pair, verifies them with a hybrid critic, ranks them using video-claim support scores, and finally consolidates them into a citation-backed report.

3.1 Evidence Stream

3.1.1 Preprocessing.

We preprocess long source videos by splitting them into fixed-size chunks of at most 120 seconds using PyAV. This prevents the VLM from silently truncating long videos under a fixed frame budget and allows each segment to be processed without exceeding memory or context constraints. We retain a mapping from each chunk identifier to its parent video identifier, and use this mapping to restore parent video IDs and consolidate the outputs.

3.1.2 Per-video ASR and translation.

Each unique video is transcribed once and cached for reuse. We use Qwen3-ASR-1.7B (Shi et al., 2026) as the primary ASR backend. For languages outside its supported set in our data, such as Burmese and Nepali, we fall back to Whisper-large-v3 (Radford et al., 2022). For non-English videos,

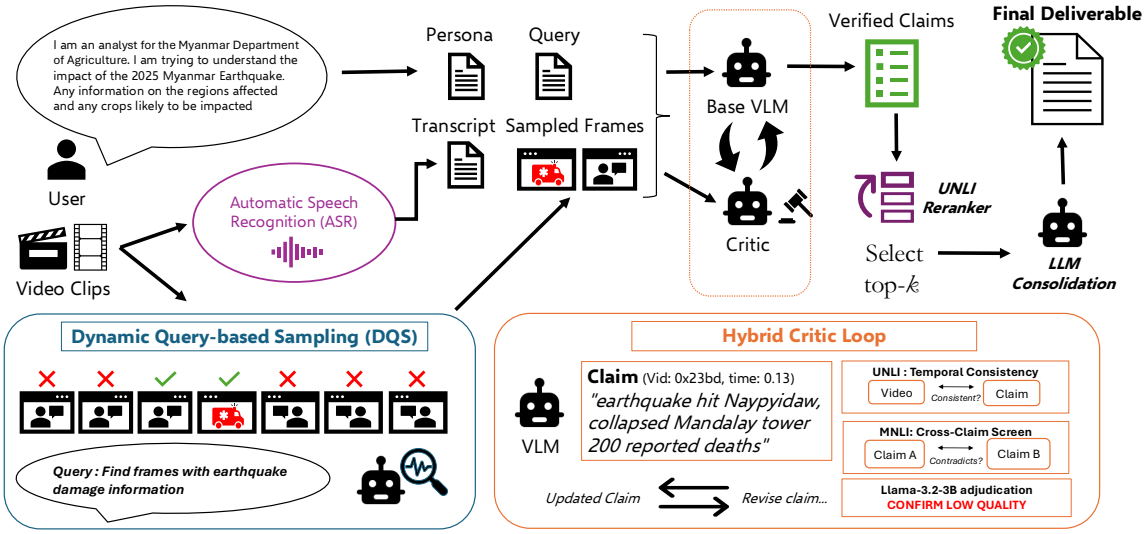


Figure 1: **Overview of CRAFT.** Given a persona, query, and relevant videos, CRAFT builds a query-specific multimodal evidence stream: each video is transcribed once via ASR, and *Dynamic Keyframe Selection* (DKS) selects the frames most relevant to the query. The base VLM consumes the persona, query, transcript, and sampled frames to produce atomic claims, which are refined by a *hybrid critic loop*—UNLI for temporal grounding, MNLI for cross-claim contradiction screening, and a Llama-3.2-3B adjudicator that confirms low-quality claims and returns repair feedback for re-extraction. Verified claims are UNLI-reranked, and the top-k are consolidated by an LLM into a report with every statement traceable to its source video and timestamp.

we also run a translation pass to obtain an English transcript. During claim extraction, we provide both the original transcript and the English translation to the VLM, allowing the model to ground claims in spoken content as well as visual evidence. Because ASR systems can produce repetitive token loops on low-resource or noisy audio (Koenecke et al., 2024), we filter degenerate transcripts before they reach the VLM. We flag a transcript as unreliable if it contains at least 20 tokens and has very low lexical diversity, measured by a type-token ratio below 0.18, where the type-token ratio is the number of unique tokens divided by the total number of tokens. We also flag transcripts with obvious local repetition, such as the same token appearing at least 8 times consecutively, or phrase-level repetition, where a single 3-token phrase accounts for at least 40% of all 3-token phrases in the transcript. Flagged transcripts are excluded from the prompt to avoid propagating ASR artifacts into downstream claims. Although this filtering may discard useful information from resource-scarce-language videos, stronger multilingual ASR systems could mitigate this limitation, which we leave for future work.

3.1.3 Dynamic Keyframe Selection.

Long videos contain many frames that are irrelevant to a given query, and uniform frame sampling can dilute the visual evidence passed to the VLM. We therefore use Dynamic Keyframe Selection (DKS) to construct a compact visual input for each query-video pair. DKS is applied independently for each pair (q, v), so the same video may yield different selected frames for different queries.

For a query q and video v, we first sample candidate frames at a fixed temporal rate. Each frame is embedded with a visual encoder and scored against the query embedding using image-text similarity:

$$s_i = \mathrm{sim}(\phi_I(f_i), \phi_T(q)),$$

where ϕ_I and ϕ_T denote the image and text encoders. The resulting scores form a query-conditioned relevance curve over the video. We use CLIP (Radford et al., 2021) Image and Text encoders.

We then select frames that balance high relevance with temporal coverage, similar to (Tang et al., 2025). The selected frame indices are sorted in temporal order and re-encoded as a short query-specific clip. During claim extraction, the resolver first checks whether a DKS clip exists for the cur-

rent query-video pair. If available, the VLM receives this compact clip instead of the full chunked video; otherwise, the pipeline falls back to the original chunk. Thus, DKS focuses visual input on query-relevant evidence while remaining optional and non-blocking.

3.2 Query-Conditioned Claim Extraction

Given the evidence stream for a query-video pair, we extract a set of source-grounded claims. For each query q and each video $v \in \mathcal{V}_q$ associated with that query, we issue one VLM call to Qwen3.5-9B (Qwen Team, 2026) served with vLLM (Kwon et al., 2023). The prompt contains the persona title, persona background, query text, the resolved video input from the evidence stream, and the cached ASR transcript when available. If persona title/background is not available based on the query and claims we use LLM (Qwen Team, 2026) to generate it in preprocessing. The model is instructed to output *atomic* claims, where each claim is a single declarative statement that can be judged as supported or unsupported by the source video.

This produces an initial per-video claim set $\mathcal{C}^0_{q,v}$ for each query-video pair. Claim extraction is performed independently for each video so that every claim remains tied to a specific source video, timestamp, and evidence modality.

Atomic claim format. Each extracted claim must be independently verifiable. We discourage compound claims that combine multiple events, entities, or causal relations into a single sentence, since such claims become unsupported if any subclause is not grounded in the video. Each claim is also tagged with its evidence modality, such as visual evidence, on-screen text, transcript, or ASR-derived speech.

3.3 Critic-Guided Claim Refinement

The initial VLM extraction can still produce claims with weak visual grounding, incorrect temporal references, or contradictions. To reduce these errors, we apply a critic-guided refinement loop separately to each query-video claim set $\mathcal{C}^0_{q,v}$. The loop runs for up to $R = 4$ rounds and combines three complementary critics.

The critic loop targets three distinct error types. First, a UNLI-based video-claim entailment model (Chen et al., 2020) checks temporal grounding by scoring each claim against its cited video segment. Claims scoring below 0.05 are marked as unsupported at the cited timestamp and ignored, while scores in $[0.05, 0.5)$ are treated as weak support and warrant re-extraction. This filters claims that may be plausible but are not grounded in the selected temporal window.

Second, a DeBERTa-v3 MNLI cross-encoder (He et al., 2023) screens the per-video claim set for possible contradictions. For each pair of claim texts, the cross-encoder estimates entailment, neutrality, and contradiction probabilities. Pairs whose contradiction probability exceeds a low threshold of 0.5 are retained as candidates for further stage. We use this stage as a high-recall filter rather than a final judge, since text-only NLI can produce false positives for claims that mention related but compatible facts.

Third, a Llama-3.2-3B adjudicator (Meta AI, 2024) verifies the candidate contradictions. Given the two claims and the MNLI contradiction score, it decides whether the claims are genuinely inconsistent spitting binary output and, if it is inconsistent, it also returns an explanation and a repair hint. The critic report is then fed back to the VLM together with the previous claim set, and the VLM re-extracts a revised set of claims by removing unsupported statements, correcting weakly grounded claims, or resolving contradictions. The loop terminates early when the claim set no longer changes. We denote the final refined per-video claim set as $\mathcal{C}_{q,v}$.

3.4 Query-Level Evidence Pooling

After per-video refinement, we aggregate claims across all videos associated with the same query. For a query q, the refined claims from each relevant video $v \in \mathcal{V}_q$ are concatenated into a query-level evidence pool:

$$\mathcal{P}_q = \biguplus_{v \in \mathcal{V}_q} \mathcal{C}_{q,v}.$$

Here, $\uplus$ denotes concatenation of claim records, not semantic deduplication. Each record remains associated with its source video, timestamp, modality, and claim identifier. This preserves provenance when the same fact is supported by multiple videos: overlapping claims are retained as distinct evidence items at this stage, and redundancy is resolved only during final inference by emitting the shared fact once with all supporting citations.

3.5 Claim Scoring and Calibration

Every refined claim in the query-level evidence pool is rescored against its source video using the same UNLI model used by the critic. This produces a support confidence score in $[0, 1]$ for each claim. We use these scores to rank evidence rather than apply a hard threshold, since thresholding can remove rare but useful evidence from long-tail videos.

For each query, the top-ranked claims form a compact claim packet for downstream inference. This packet keeps the strongest supported evidence while retaining source identifiers required for citation generation.

3.6 Citation-Preserving Inference

The final inference stage uses Qwen3.5-9B in text-only mode to convert the calibrated claim packet into report statements. The model is constrained to use only information present in the packet and to avoid adding new entities, numbers, dates, or causal links.

Redundant evidence is handled by citation merging: when multiple claims support the same fact, the report states the fact once and attaches all corresponding source identifiers. This preserves citation coverage without repeating semantically identical statements. Final report sections are populated directly from the generated inferences and their associated source identifiers; during submission formatting, chunk-level video IDs are remapped to their parent video IDs before writing the JSONL file.

4 Experiments

4.1 Benchmarks

MAGMaR. We evaluate on the MAGMaR 2026 oracle task, a multi-video question answering benchmark targeting real-world news events. The data is based on subset of WikiVideo (Martin et al., 2025a). For the retrieval and RAG settings, we retrieve relevant videos from a combination of the MAGMaR data and MultiVENT2.0 test (Kriz et al., 2025). The dataset comprises 92 source videos with average length of 1.82 mins distributed across 10 topically diverse topics including elections, natural disasters, and geopolitical events paired with 19 official evaluation queries. Each query is associated with a set of relevant videos, and the answer may require aggregating evidence distributed across multiple clips. This multi-source setting makes the benchmark challenging because models must identify relevant evidence across heterogeneous videos while avoiding unsupported claims from irrelevant or redundant content. Each generated claim should also be accompanied by a citation to the supporting evidence video.

WikiVideo. We also evaluate on the original super set dataset - WikiVideo (Martin et al., 2025a), a grounded multi-video article generation benchmark built from real-world event videos linked to Wikipedia articles. The dataset is constructed from MultiVENT 1.0 and 2.0 (Kriz et al., 2025) videos whose events have corresponding English Wikipedia articles, and the reference articles are derived from Wikipedia lead sections. WikiVideo contains 57 event topics spanning 427 videos with average length of 1 min from 2016 to 2025, with each event paired with an expert-written Wikipedia-style article grounded in video evidence. The annotation process decomposes Wikipedia lead sentences into atomic claims, grounds each claim in supporting video, audio, or OCR evidence, and rewrites the article so that it includes only information supported by the videos. On average, each event contains 7.65 relevant videos, 51.1 grounded subclaims, and a 118-token reference article. This makes WikiVideo well suited for evaluating whether models can synthesize high-level event information across multiple videos while maintaining claim-level grounding and citations to supporting evidence.

4.2 Evaluation Metrics

Predictions are evaluated using both automatic and human evaluation. For automatic evaluation, we use MiRAGE (Martin et al., 2025b), which assesses factuality, information coverage, groundedness, and the correctness of citation attribution. Each MiRAGE entailment judgment is judged by Qwen-7B or CLUE (Zhang et al., 2026). Reported results in the main text use Qwen-7B, which was used during the development of our CRAFT system for submission. The official MAGMaR leaderboard uses CLUE for evaluation, we report these results in the supplementary material. For human evaluation, three annotators assign scalar scores from 1 to 5 to each system output, assessing factuality, adequacy, coherence, relevance, and fluency. After scoring all predictions, the annotators also select the best system response for each query. We report the human evaluation results in the supplementary

System	MAGMaR-Test							WikiVideo						
	Ref-P	Ref-R	Ref-F1	Cite-P	Cite-R	Cite-F1	Avg	Ref-P	Ref-R	Ref-F1	Cite-P	Cite-R	Cite-F1	Avg
Molmo2-8B	0.623	0.541	0.579	0.498	0.421	0.457	0.518	0.641	0.682	0.661	0.512	0.598	0.552	0.607
InternVL-3.5-30B-A3B	0.749	0.688	0.717	0.645	0.521	0.576	0.649	0.802	0.821	0.811	0.731	0.689	0.710	0.761
(+ ASR)	0.761	0.722	0.741	0.659	0.551	0.600	0.672	0.815	0.848	0.831	0.743	0.712	0.727	0.779
Gemma-4-31B	0.701	0.658	0.679	0.589	0.532	0.559	0.620	0.721	0.748	0.734	0.618	0.630	0.624	0.679
(+ ASR)	0.712	0.701	0.706	0.601	**0.561**	0.580	0.644	0.732	0.778	0.754	0.629	0.651	0.640	0.697
CRAFT Baseline	0.437	0.756	0.430	0.875	0.251	0.359	0.518	0.833	0.834	0.834	0.951	0.662	0.764	0.814
+ Critic Loop	0.491	0.766	0.480	0.854	0.259	0.360	0.535	0.859	0.845	0.842	0.953	0.668	0.773	0.822
+ Atomic Claims	0.808	0.762	0.764	**0.944**	0.336	0.426	0.673	0.940	0.620	0.735	0.855	**0.858**	**0.848**	0.809
+ ASR	0.760	**0.810**	**0.783**	0.935	0.512	**0.635**	**0.739**	0.871	**0.849**	**0.854**	0.949	0.656	0.762	0.823
↓ frames (uniform)	0.775	0.775	0.769	0.902	0.503	0.616	0.723	0.930	0.640	0.746	0.845	0.844	0.830	0.805
↓ frames (DKS)	**0.822**	0.743	0.772	0.927	0.453	0.574	0.715	**0.940**	0.832	0.797	**0.966**	0.647	0.761	**0.824**

Table 1: **Main results on MAGMaR-Test and WikiVideo**. Baseline VLMs are evaluated both with and without ASR transcript access. All rows of CRAFT baseline for MAGMaR-Test use Qwen3.5-9B and Qwen3-VL-30B-Instruct for WikiVideo as the base VLM. Best results per column are **bolded**. Avg denotes the mean of all six metrics. We use 128 uniformly sampled frames except last two rows. ↓ denotes a reduced-frame setting used to stress test uniform sampling; DKS improves this setting by selecting more query-relevant frames, especially improving precision. For MAGMaR-Test we choose 64 reduced frames and for WikiVideo we choose 32 reduced frames.

System	ROUGE-L	BERTScore	AnsRel
MAGMaR-Test			
InternVL-3.5-30B-A3B	0.1497	0.0945	0.6382
(+ ASR)	0.1182	0.0964	0.6462
Gemma-4-31B	0.1426	0.0950	0.5769
(+ ASR)	0.1100	0.1224	0.5799
CRAFT	**0.1839**	**0.1709**	**0.6504**
WikiVideo			
InternVL-3.5-30B-A3B	0.1241	-0.0184	0.5843
(+ ASR)	0.1265	0.0083	0.6069
Gemma-4-31B	0.1526	0.0634	0.6486
(+ ASR)	0.1360	0.0632	0.6589
CRAFT	**0.3014**	**0.2683**	**0.6664**

Table 2: **Generation quality comparison on MAGMaR-Test and WikiVideo.** We report ROUGE-L, BERTScore F1, and Answer Relevance (AnsRel) for baseline VLMs with and without ASR transcript access, alongside CRAFT.

material.

Concretely, we report six MiRAGE (Martin et al., 2025b) metrics that evaluate both information quality and citation fidelity at the subclaim level. *Reference Precision (Ref-P)* measures the proportion of generated subclaims that are supported by the reference, capturing whether the prediction contains factual and relevant information. *Reference Recall (Ref-R)* measures the proportion of reference subclaims that are covered by the generated report, capturing information completeness. Their harmonic mean gives *Reference F1 (Ref-F1)*. For citation evaluation, *Citation Precision (Cite-P)* measures whether generated subclaims are supported by their cited source videos, while *Citation Recall (Cite-R)* measures whether reference subclaims that are covered by the prediction are attributed to the correct supporting videos. Their harmonic mean gives *Citation F1 (Cite-F1)*. The overall *Macro-Average* is computed as the mean of the six reported metrics. Additionally, we report three complementary metrics designed to capture failure modes not explicitly measured by MiRAGE:

ROUGE-L (Lin, 2004), computed over the concatenated report text without stemming. Since the benchmark spans multiple languages (e.g., English, Mandarin, Burmese, and Nepali), language-specific stemming introduces substantial noise. We therefore use ROUGE-L primarily as a lightweight regression signal for lexical overlap with the reference report.

BERTScore (Zhang et al., 2020) F1 using `bert-base-multilingual-cased` with `rescale_with_baseline=True`. This metric captures document-level semantic similarity and stylistic alignment, complementing MiRAGE's claim-level decomposition.

RAGAS Answer Relevance (Es et al., 2024), which directly evaluates whether the persona-grounded query was meaningfully answered. For each generated report, we sample $K = 3$ hypothetical questions using `Qwen2.5-7B-Instruct` (temperature 0.7, top-p = 0.9), embed both the reconstructed and gold queries using `Qwen3-Embedding-0.6B`, and report the mean co-

sine similarity.

ROUGE-L and BERTScore are reference-dependent metrics and are therefore computed only on the subsets containing gold reports (8/19 queries for MagMaR and 52/56 queries for WikiVideo). The remaining 15 queries are excluded from these metrics and explicitly marked in the results table. In contrast, Answer Relevance is reference-free and is reported for all queries.

4.3 Baselines and Setup

CRAFT Baseline. We construct the CRAFT baseline as a basic pipeline for generating answers and citations given a video and its corresponding query. Additional proposed improvements are built on top of this baseline. The pipeline uses a multimodal LLM (base VLM) as the backbone: for each query, the model receives sampled video frames and is prompted to generate claims along with their supporting video citations. The model only has access to frames that are uniformly sampled from the input video, with a maximum of 128 frames provided. In the baseline, we also use UNLI (Chen et al., 2020) to re-rank the generated claims so that the downstream LLM can better prioritize important evidence. Finally, a text-only LLM aggregates the claims, removes duplicates, and consolidates them into the final response for each query. CRAFT uses base VLM as Qwen-3.5-9B-VL as a backbone for MAGMaR-Test benchmark and Qwen3-VL-30B-A3B-Instruct for Wikivideo benchmark, unless otherwise explicitly specified. For final LLM Consolidator we use Qwen3.5-9B in text-only mode. Every other addition over this baseline is described in section 3 and evaluated in table 1. Results for CRAFT are obtained using 8 NVIDIA A6000 GPUs, and it takes 2 hours to get final results for Wikivideo and 0.75 hour on MAGMaR-Test dataset.

Other Baselines. We additionally evaluate a diverse set of publicly available multimodal LLMs spanning multiple architectural families and parameter scales, including Molmo2-8B (Clark et al., 2026), InternVL3-30B-A3B (Zhu et al., 2025), Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025a), and Gemma-4-31B (Team et al., 2024). These comparisons provide a broader characterization of the proposed task beyond the CRAFT pipeline itself.

For all baselines, videos are represented using uniformly sampled frames. For InternVL3-30B-A3B and Gemma-4-31B, we further evaluate both *visual-only* and *visual+ASR* variants using the same ASR backend employed by CRAFT. Concretely, for each (q, v) pair, we issue a single VLM call requesting factual claims relevant to the query and concatenate the resulting per-video generations into a final per-query report without any additional scoring, reranking, deduplication, or calibration.

Long videos are pre-segmented offline into 60-second chunks. Each chunk is sampled at 1 fps with a maximum of 60 frames per call, and generation is capped at 1024 new tokens.

In the *visual+ASR* setting, we augment the visual inputs with `Whisper-large-v3` transcripts sourced from the `akhilvssg/magmar-2026-test-asr-embeddings` release on MagMaR and the corresponding WikiVideo dump. For each chunk, we provide both the original-language transcript and its English translation as auxiliary textual context.

4.4 Main Results

Table 1 reports the main results on MAGMaR-Test and WikiVideo. Overall, CRAFT achieves the best average performance on MAGMaR-Test and competitive performance on WikiVideo, showing consistent gains over publicly available VLM baselines. Among the baseline models, adding ASR generally improves performance, especially on WikiVideo, indicating that explicit speech transcripts provide useful evidence beyond visual frames alone.

Within CRAFT, the largest improvement comes from moving beyond the initial baseline toward atomic claim generation and ASR-augmented evidence extraction. On MAGMaR-Test, adding atomic claims substantially improves Ref-P and Ref-F1 as compared to baseline, suggesting that decomposing evidence into finer-grained claims helps the model produce more precise and verifiable answers. Adding ASR further improves Ref-R and Cite-F1, showing that spoken content is important for recovering missing information and assigning better citations. However, citation recall remains more challenging than citation precision, indicating that exact claim-to-video attribution is still a difficult part of the task.

The last two rows simulate low-frame settings to stress test the robustness of frame sampling when only small compute budget is alloted to the task. This becomes more challenging for longer videos, where relevant information is often sparse and distributed across distant segments, making it harder to preserve context. This is reflected

Variant	Ref-P	Ref-R	Ref-F1	Cite-P	Cite-R	Cite-F1	Avg
Qwen3.5-9B-VL backbone	0.760	0.810	0.783	0.935	0.512	0.635	0.739
Qwen3-Omni-30B-A3B	0.745	0.761	0.735	0.878	0.346	0.471	0.656

Table 3: **Backbone replacement** ablation on MAGMaR-Test. Qwen3-Omni-30B-A3B directly uses audio input, while Qwen3.5-9B-VL uses ASR transcripts. Avg denotes the mean of all six metrics.

in the larger performance drop across most metrics on MAGMaR-Test compared to WikiVideo, as MAGMaR-Test videos are on average roughly twice as long. The ↓ frames rows denote reduced-frame settings, where fewer frames are passed to the system. In the uniform setting, the reduced frame budget is sampled uniformly, which can miss query-relevant evidence. In the DKS setting, uniform sampling is replaced with dynamic keyframe selection under the same reduced-frame budget. DKS improves precision in several cases by selecting more relevant frames, although it can trade off recall when some supporting evidence is filtered out. This suggests that adaptive frame selection is useful under constrained visual budgets, but further work is needed to balance precision-oriented keyframe selection with broad evidence coverage.

Table 2 reports auxiliary generation-quality metrics on MAGMaR-Test and WikiVideo. CRAFT achieves the best ROUGE-L, BERTScore F1, and Answer Relevance on both datasets, indicating that its claim-centric aggregation improves not only factual grounding and citation quality, but also the fluency and relevance of the generated reports. For the baseline VLMs, adding ASR generally improves answer relevance and semantic similarity in some cases, but the gains are not consistent across all metrics.

4.5 Ablation Studies

Omni-Model. Although Qwen3-Omni-30B-A3B directly processes audio, it does not outperform the ASR-based Qwen3.5-9B-VL backbone as seen in table 3. This suggests that, for claim-centric video QA, explicit ASR transcripts provide a more reliable intermediate representation for evidence extraction, citation assignment, and downstream text-based verification. Direct audio conditioning may encode speech information implicitly, but it can make fine-grained details such as named entities, dates, and numerical facts harder to recover and verify. In contrast, ASR converts speech into explicit textual evidence,

System	Ref-P	Ref-R	Ref-F1	Cite-P	Cite-R	Cite-F1	Avg
CRAFT (full)	0.760	0.810	0.783	0.935	0.512	0.635	0.739
w/ Qwen replaces UNLI	0.732	0.788	0.759	0.874	0.469	0.601	0.704
w/ Qwen replaces Llama-3.2-3B	0.763	0.812	0.787	0.937	0.516	0.619	0.732
w/ Qwen unified critic (no MNLI screen)	0.743	0.798	0.770	0.909	0.493	0.619	0.722

Table 4: **Component ablations on MAGMaR-Test.** Replacing specialized critic components with a unified Qwen-based adjudicator consistently degrades attribution performance. The unified critic variant removes the DeBERTa-v3 MNLI screening stage and performs contradiction detection and adjudication in a single pass. We report precision (P), recall (R), and F1 for both Reference Attribution and Citation Attribution.

which better aligns with the claim aggregation and citation modules in CRAFT.

UNLI Scorer. Replacing UNLI with zero-shot Qwen3.5-9B causes the largest drop in citation metrics, confirming that UNLI's specialized temporal entailment training is not recoverable by a general-purpose VLM.

Critic Adjudicator. Replacing Llama-3.2-3B with Qwen3.5-9B yields a marginal drop, suggesting the 3B adjudicator is already sufficient for binary contradiction confirmation and the larger model provides no measurable benefit.

Unified Qwen Critic. Removing the DeBERTa MNLI pre-filter and collapsing screening and adjudication into a single Qwen pass degrades citation precision, showing the specialized NLI screener provides a signal that general-purpose prompting does not fully replicate.

5 Conclusion and Future Work

We presented CRAFT, a claim-centric pipeline for grounded multi-video question answering that combines keyframe selection, ASR-based evidence extraction, critic-guided verification, and citation-backed report generation. CRAFT improves over the baseline through atomic-claim formatting, ASR, and the critic loop. However, recall and citation recall remain challenging, suggesting that future work should improve evidence coverage, cross-video retrieval, multilingual ASR, and precise claim-to-video attribution.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection.

In *The Twelfth International Conference on Learning Representations (ICLR)*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 5 others. 2025a. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online. Association for Computational Linguistics.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, and 1 others. 2026. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*.

Jisheng Dang, Huilin Song, Junbin Xiao, Bimei Wang, Han Peng, Haoxuan Li, Xun Yang, Meng Wang, and Tat-Seng Chua. 2025. MUPA: Towards multi-path agentic reasoning for grounded video question answering. *arXiv preprint arXiv:2506.18071*.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, and 1 others. 2025. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 91–104.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*.

Hong Gao, Yiming Wang, Xin Hu, Xun Cao, and Mingkui Tao. 2025. APVR: Hour-level long video understanding with adaptive pivot visual information retrieval. *arXiv preprint arXiv:2506.04953*. To appear in AAAI 2026.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. MA-LMM: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations (ICLR)*.

Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. 2025. VideoRAG: Retrieval-augmented generation over video corpus. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21278–21298.

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, pages 1672–1681.

Reno Kriz, Kate Sanders, David Etter, Kenton Murray, Cameron Carpenter, Kelly Van Ochten, Hannah Recknor, Jimena Guallar-Blasco, Alexander Martin, Ronald Colaianni, Nolan King, Eugene Yang, and Benjamin Van Durme. 2025. MultiVENT 2.0: A massive multilingual benchmark for event-centric video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. 2025. VidHalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13723–13733.

Jianxin Liang, Xiaojun Meng, Yueqian Wang, Chang Liu, Qun Liu, and Dongyan Zhao. 2024. End-to-end video question answering with frame scoring mechanisms and adaptive sampling. *arXiv preprint arXiv:2407.15047*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. 2026. VideoMind: A chain-of-LoRA agent for temporal-grounded video reasoning. In *The Fourteenth International Conference on Learning Representations (ICLR)*.

Alexander Martin, Reno Kriz, William Gantt Walden, Kate Sanders, Hannah Recknor, Eugene Yang, Francis Ferraro, and Benjamin Van Durme. 2025a. Wikivideo: Article generation from multiple videos. *arXiv preprint arXiv:2504.00939*.

Alexander Martin, William Walden, Reno Kriz, Dengjia Zhang, Kate Sanders, Eugene Yang, Chihsheng Jin, and Benjamin Van Durme. 2025b. Seeing through the mirage: Evaluating multimodal retrieval augmented generation. *arXiv preprint arXiv:2510.24870*.

Meta AI. 2024. Llama 3.2: 1b and 3b instruct model card.

Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. 2024. MoReVQA: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13235–13245.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. 2025. VideoRAG: Retrieval-augmented generation with extreme long-context videos. *arXiv preprint arXiv:2502.01549*.

Saron Samuel, Dan DeGenaro, Jimena Guallar-Blasco, Kate Sanders, Oluwaseun Eisape, Tanner Spendlove, Arun Reddy, Alexander Martin, Andrew Yates, Eugene Yang, Cameron Carpenter, David Etter, Efsun Kayi, Matthew Wiesner, Kenton Murray, and Reno Kriz. 2025. MMMORRF: Multimodal multilingual modularized reciprocal rank fusion. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. 2024. LongVU: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, and 1 others. 2026. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*.

Yan Shu, Peitian Zhang, Zheng Liu, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. 2025. Video-XL: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485*.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. 2024. MovieChat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18221–18232.

Guangyu Sun, Archit Singhal, Burak Uzkent, Mubarak Shah, Chen Chen, and Garin Kessler. 2025a. From frames to clips: Efficient key clip selection for long-form video understanding. *arXiv preprint arXiv:2510.02262*.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. 2025b. MDP^3: A training-free approach for list-wise frame selection in video-LLMs. *arXiv preprint arXiv:2501.02885*. Published at ICCV 2025.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. 2025. Adaptive keyframe sampling for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29118–29128.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2025. Correctness is not faithfulness in retrieval augmented generation attributions. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pages 22–32.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024a. VideoAgent: Long-form video understanding with large language model as agent. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. 2024b. VideoHallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2025. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3272–3283.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, Jiajun Wu, and Manling Li. 2025. Rethinking temporal search for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8579–8591.

Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Ji-Rong Wen, and Zhicheng Dou. 2025. VideoDeepResearch: Long video understanding with agentic tool using. *arXiv preprint arXiv:2506.10821*.

Nianbo Zeng, Haowen Hou, Fei Richard Yu, Si Shi, and Ying Tiffany He. 2025. SceneRAG: Scene-level retrieval-augmented generation for video understanding. *arXiv preprint arXiv:2506.07600*.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2024a. A simple LLM framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 21715–21737.

Dengjia Zhang, Alexander Martin, William Jurayj, Kenton Murray, Benjamin Van Durme, and Reno Kriz. 2026. Unified multimodal uncertain inference. *arXiv preprint arXiv:2604.08701*.

Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, Zhiyu Tan, Hao Li, Xingjun Ma, and Jingjing Chen. 2024b. EventHallusion: Diagnosing event hallucinations in video LLMs. *arXiv preprint arXiv:2409.16597*.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. 2025a. Q-Frame: Query-aware frame selection and multi-resolution adaptation for video-LLMs. *arXiv preprint arXiv:2506.22139*. Published at ICCV 2025.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.

Xian Zhang, Zexi Wu, Zinuo Li, Hongming Xu, Luqi Gong, Farid Boussaid, Naoufel Werghi, and Mohammed Bennamoun. 2025b. AdaRD-Key: Adaptive relevance–diversity keyframe sampling for long-form video understanding. *arXiv preprint arXiv:2510.02778*.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. 2025c. Deep video discovery: Agentic search with tool use for long-form video understanding. *arXiv preprint arXiv:2505.18079*.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024c. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.

Zhuo Zhi, Qiangqiang Wu, Minghe Shen, Wenbo Li, Yinchuan Li, Kun Shao, and Kaiwen Zhou. 2025. VideoAgent2: Enhancing the LLM-based agent system for long-form video understanding by uncertainty-aware CoT. *arXiv preprint arXiv:2504.04471*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, and 2 others. 2025. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Yuanhao Zou, Yifan Liu, Yang Liu, Yifan Zhang, Han Zhang, and Chen Chen. 2025. A.I.R.: Enabling adaptive, iterative, and reasoning-based frame selection for video question answering. *arXiv preprint arXiv:2510.04428*.

Appendix

A MiRAGE Results using CLUE

Table 5 reports per-query MiRAGE scores using CLUE as the evaluation backbone. The results show that CRAFT obtains stronger information precision than recall, indicating that its generated claims are often relevant and supported, but do not cover all reference subclaims. This is expected because CRAFT is designed to be conservative: it filters, deduplicates, and consolidates evidence to avoid unsupported statements, which improves factuality but can reduce coverage.

Citation scores are lower, especially citation recall. In MiRAGE, citation precision measures whether generated subclaims are supported by their cited videos, while citation recall measures whether the covered reference information is attributed to the correct supporting videos. This makes citation recall particularly challenging in MAGMaR, where evidence may be distributed across multiple heterogeneous videos and several videos may contain overlapping or partial support for the same event. As a result, a prediction can contain correct information but still lose citation recall if the exact supporting video is missing, incomplete, or not aligned with the evaluator's expected grounding. These results are therefore consistent with the main-text findings: CRAFT is relatively effective at producing factual content, but exact claim-to-video attribution remains the harder part of the task.

Topic	Info F1		Cite F1	
	P	R	P	R
Average*	**72.4**	**36.1**	**60.5**	**24.2**
2025_Myanmar_earthquake_q1	72.7	86.7	59.1	80.0
Liberation_Day_Tariffs_q1	70.0	64.1	65.0	66.7
Blue_Ghost_Mission_1_q2	70.8	57.1	70.8	39.3
Shi_Yongxin_Scandal_q1	76.7	40.2	86.7	31.1
Shi_Yongxin_Scandal_q2	95.2	36.4	95.2	30.3
Blue_Ghost_Mission_1_q1	76.5	39.3	64.7	35.7
Liberation_Day_Tariffs_q2	70.6	41.0	70.6	30.8
2025_Alaskan_Typhoon_q2	88.9	36.5	0.0	0.0
Nepal_Youth_Protests_q2	92.9	29.4	96.4	23.5
2025_Alaskan_Typhoon_q1	72.7	28.6	4.5	0.0
Tropical_Storm_Wipha_q1	96.6	20.3	96.6	7.1
2025_Canadian_federal_election_q2	28.6	30.6	35.7	11.1
Nepal_Youth_Protests_q1	63.6	13.2	63.6	4.4
Palisades_Fire_q2	100.0	9.5	100.0	2.6
Palisades_Fire_q1	78.3	7.9	47.8	2.1
2025_Canadian_federal_election_q1	3.6	36.1	10.7	22.2

Table 5: Per-topic CLUE reference scores for the CRAFT submission. Info F1 and Cite F1 are reported with precision (P) and recall (R). *We exclude queries with missing source videos from the MAGMaR-Test average, as these cases produce flat zero scores independent of system quality.

B Human Evaluation

The human evaluation, as shown in table 6, indicate that CRAFT produces reasonably useful responses in several cases, but it is not yet consistently preferred over competing systems on MAGMaR leader-board. These results suggest that future improvements should focus on increasing information coverage and strengthening claim-to-video citation alignment, while preserving CRAFT's emphasis on grounded and conservative generation.

C Pre-Processing Details of WikiVideo

To evaluate WikiVideo using the same structure as the MAGMaR test set, we convert the WikiVideo annotations into a MAGMaR-style format. We start with 56 candidate WikiVideo events and remove four events that overlap with the MAGMaR 2026 evaluation set, resulting in 52 events. For each event, we keep only reference claims that are supported by at least one video, and we further retain only events with at least three video-supported claims.

For each remaining event, an LLM agent generates a triplet <persona_title, background, query> following the MAGMaR persona-query format. We then perform an audit step in which each generated triplet is scored on 5-point criteria on following axis: persona lignment, query answerability given article, and overall grounding. Items that are flagged during this audit are rewritten and rescored. The final dataset includes only events with an overall grounding score of at least 4, yielding a 52-query WikiVideo evaluation set.

Finally, the audited persona_title, background, and query triples are converted into MAGMaR-style query, ground-truth, and topic-video files, allowing WikiVideo to be evaluated directly with the same pipeline used for MAGMaR.

Topic	Query	Avg. Score	Best Votes
Overall	–	**2.542**	**0 / 57**
2025-Alaska-typhoon	q1	3.000	0 / 3
2025-Alaska-typhoon	q2	2.667	0 / 3
2025_Myanmar_earthquake	q2	3.000	0 / 3
2025_Palisades_fires	q1	2.667	0 / 3
2025_Palisades_fires	q2	2.000	0 / 3
2025_Shi_Yongxin_Scandal	q1	3.000	0 / 3
2025_Shi_Yongxin_Scandal	q2	2.333	0 / 3
2025_Tropical_Storm_Wipha	q2	2.333	0 / 3
2025_canadian_federal_election	q1	2.000	0 / 3
2025_canadian_federal_election	q2	1.667	0 / 3
2025_nepal_youth_protests	q1	2.667	0 / 3
2025_nepal_youth_protests	q2	2.667	0 / 3
Blue_Mission_Ghost_1	q1	2.333	0 / 3
Blue_Mission_Ghost_1	q2	2.333	0 / 3
Liberation-Day-tariffs	q1	3.333	0 / 3
Liberation-Day-tariffs	q2	2.667	0 / 3

Table 6: Official human evaluation results for the CRAFT submission. Avg. Score is the mean scalar score on a 1–5 scale. Best Votes denotes the number of annotators who selected our system as the best response for the query. The overall scalar score is 2.542 with standard deviation 0.676 over 48 annotations.

Findings of the MAGMaR 2026 Shared Task

Alexander Martin[1] **Dengjia Zhang**[1] **Joel Brogan**[2*] **Francis Ferraro**[3]
Jeremy Gwinnup[4] **Reno Kriz**[5] **Teng Long**[6]
Kenton Murray[5] **Andrew Yates**[5] **Xiang Xiang**[7]
[1]Johns Hopkins University [2]OpenAI
[3]University of Maryland, Baltimore County [4]Air Force Research Laboratory
[5]Human Language Technology Center of Excellence, Johns Hopkins University
[6]University of Amsterdam [7]Huazhong University of Science and Technology
{amart233, kenton, rkriz1}@jhu.edu

Abstract

This overview paper presents the results of the shared task for the second workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR). In this shared task participants submitted systems focused on either (i) video retrieval or (ii) grounded generation of articles given retrieved videos. Teams could submit to either task. For the retrieval task, we had 2 participating teams that submitted a total of 17 systems – all of which beat a baseline derived from the winner of last year's shared task. On the generation side, we had 4 teams submit 16 systems. All teams had at least one generated report that was labeled the best by a human annotator.

1 Introduction

The second workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2026) was located at ACL in San Diego and hosted a shared task focused on video retrieval and generation. Increasingly, information seeking needs require solutions that span modalities, yet much of the literature, and open shared tasks, continue to focus on one modality (i.e., text). The focus of this shared task is aimed at addressing this deficiency. The task built upon the success of last year's shared task of retrieving videos given a query. This year the task expanded the test collection of videos and introduced a new generation track:

- **Retrieval Track:** Systems provide a ranked list of videos in the collection ordered by relevance to the query.
- **Generation Track:** Systems produce a text article that answers the information need and grounds its content in the retrieved videos.

This shared task focuses on retrieving relevant videos and generating grounded articles that respond to information needs. Given a query describing a real-world current event, participating systems must identify pertinent videos from a large multilingual, multimodal collection and use that evidence to produce a coherent and informative written article. Our retrieval collection comprised over 110,000 multilingual, event-centric videos. The generation task had 19 queries for which systems needed to produce an article that met the information seeking needs of a specific persona. The summaries were judged by 3 human participants.

Overall, this was a very successful shared task garnering multiple submissions with all teams beating very strong baselines in both tracks. Teams made meaningful improvements and explored cutting-edge solutions to an open-problem. We are delighted with the results and suggest you read the system descriptions. That being said, here are some of the highlights across submissions:

Findings of the 2026 MAGMaR Shared Task

- Text-based reasoning dominated both tracks. Converting videos into summaries and captions rather than operating on video directly helps in both retrieval and generation.
- Complex, persona-constrained queries issued against a large video corpus do not perform well for retrieval methods using only a single dense embedding. However, more expressive retrieval architectures can do substantially better.
- The best performing retrieval systems relied on broad first-stage recall with reasoning-based reranking and significantly outperformed all baselines.
- There is a clear disconnect between human preference and automatic metrics for scoring generations. Systems that abstain when evidence is insufficient are rewarded by annota-

[*]Subsequent authors listed alphabetically

Proceedings of the 2nd Workshop on Multimodal Augmented Generation via Multimodal Retrieval, pages 144–150

tors yet penalized by the automatic metric as precision failures.

2 Data

The shared task is built on WIKIVIDEO25, which extends WIKIVIDEO (Martin et al., 2025) to address its limitations in query specificity, language coverage, video diversity, and annotation methodology. Figure 1 gives a general visualization, and Appendix A provides a detailed comparison with the original dataset.

Queries and personas. Each topic comes with a persona and a query (two per topic; see Figure 2). The persona constrains what counts as relevant evidence, so systems must reason about which parts of the videos matter for a specific professional need. Each query is written in one of two ways: *biased*, where the author had watched the videos, and *unbiased*, where the author had only seen video titles. Unbiased queries may ask for information absent from the collection, testing whether systems can acknowledge gaps rather than hallucinate.

Annotation and video diversity. Ground-truth articles are built bottom-up: annotators first write atomic claims grounded in each video's audiovisual content, then combine them into higher-level inferences. This produces reports faithful to the videos and reduces the advantage of parametric knowledge. The video collection shifts toward raw and "diet raw" footage, videos captured by bystanders without professional narration, which force models to infer events from low-level audiovisual cues instead of relying on transcripts alone. WIKIVIDEO25 also adds two Mandarin-language topics and one mixed Nepali–English topic, so systems must process non-English audio and OCR.

MultiVENT 2.0. Last year's shared task only had the retrieval track. We used the test collection of MultiVENT 2.0 (Kriz et al., 2025), a dataset of over 110,000 videos across 6 languages. This year, these videos were used as a distractor set which was combined with WIKIVIDEO25 to form the test collection. The videos in MultiVENT2.0 and WIKIVIDEO25 come from similar sources, video types, and distributions. Because MultiVENT 2.0 covers events only through 2023 and all WIKIVIDEO25 events occur in 2025, no distractor video can be a true positive for any query; relevance judgments therefore apply exclusively to the WIKIVIDEO25 portion of the collection.

3 Retrieval Task

Task. The retrieval task requires systems to rank a corpus of over 110,000 multilingual, event-centric videos by relevance to a structured query. As illustrated in Figure 2, each query comes with a detailed persona and a complex information need. Systems must return a ranked list of the top 1000 videos and are evaluated by nDCG and Recall at cutoffs of 10, 20, and 100, measured using `ir-measures` (MacAvaney et al., 2022), over the combined WIKIVIDEO25 and MultiVENT 2.0 (Kriz et al., 2025) test corpora.

3.1 Systems

We provide participants with three baselines: OmniEmbed (Ma et al., 2025), an open-source dense retriever; OmniEmbed + RankVideo (Skow et al., 2026), OmniEmbed reranked by an open-source video-native reranker; and Wholembed-v3 (Mixedbread Team, 2026), a closed-source first-stage retrieval model.

We received 17 submissions[1] from two systems: C2F-RAG (Dai et al., 2026) and MARQUIS (Chakraborty et al., 2026). Both follow a similar pipeline: broad first-stage recall followed by reasoning-based reranking, with no task-specific training. C2F-RAG indexes text proxies of the video content (global summaries and keyframe captions) with BGE-M3 (Chen et al., 2025), then applies an LLM-based cognitive reranking agent that scores each candidate on logical alignment with the query persona. MARQUIS operates on the video content directly, decomposing each query into atomic sub-queries, retrieving independently over each with OmniEmbed, fusing the resulting ranked lists, and reranking with RankVideo.

3.2 Results

Table 1 summarizes the retrieval results.

First-stage recall vs. reasoning-based reranking. The baselines span a wide performance range. OmniEmbed achieves an nDCG@10 of 0.17, which shows how difficult it is to match complex, persona-constrained queries against a large video corpus with a single embedding. Adding RankVideo as a reranker more than triples this to 0.54; even over a weak first stage, a reasoning-based second stage makes a large difference. Wholembed-v3, a closed-source late-interaction model, reaches 0.72 without

[1] Only showing the top 5 from MARQUIS

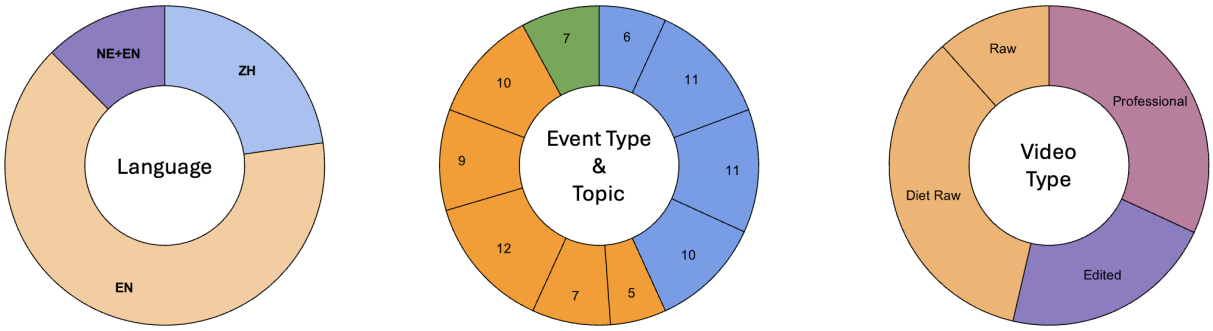


Figure 1: WIKIVIDEO25 Summary Statistics. "Language" (left): Videos are in English, Chinese, and Nepali. "Event Type & Topic" (center): are partitioned by number of videos per topic and color coded by event type. Event Type Key: Orange: Natural Disasters, Blue: Political Developments, Green: Science. "Video Type" (right): Nearly half of videos are raw and diet raw footage, captured by bystanders without professional editing or narration.

Query 1

PERSONA Statistician for North American Elections

QUERY Help me compile parliamentary and vote share statistics on the 2025 Canadian Federal elections. This should include how many seats each party won or lost, how many total seats each party now has, any information on total (popular) vote share available for the major parties and for any specific candidates for which it is available. Any demographic breakdowns of support for particular parties or candidates would also be helpful. For all statistics, please also include as much detail as possible on their source, as this helps me get a sense for how credible they are.

Query 2

PERSONA Policy Analyst at Elections Canada

QUERY I am looking to understand the key dynamics and outcomes of the 2025 Canadian Federal Election. Specifically, I am interested in how vote shares translated into seat outcomes under the first-past-the-post system, the regional patterns that enabled the Liberals to form government, and the factors contributing to the sharp decline in NDP support. Information on polling trends during the campaign, the role of leadership changes, and the influence of external political pressures—such as U.S. political rhetoric—would be particularly valuable.

Figure 2: Two example queries from WIKIVIDEO25 for the *2025 Canadian Federal Election* event. Query 1 is *unbiased* and adopts a statistician persona seeking numerical data; Query 2 is *biased* and adopts a policy analyst persona focused on structural explanations. Both target English-language output.

any reranking, so more expressive retrieval architectures can close much of the gap on their own. Both submitted systems pair broad first-stage recall with reasoning-based reranking and outperform all baselines.

Text proxies vs. video-native retrieval. C2F-RAG achieves the highest scores on every metric (nDCG@10 of 0.848) despite never operating on video content directly; it indexes text proxies (global summaries and keyframe captions) and reranks with an LLM-based cognitive agent. MARQUIS operates on video directly through OmniEmbed, yet its best configuration (MARQUIS-2) reaches only 0.759. Whether text proxies are a better retrieval substrate for complex queries than current video embeddings, or whether any reasonable first stage would perform similarly when paired with a strong reranker, remains an open question.

Query decomposition and rank fusion. MARQUIS decomposes each query into atomic subqueries, retrieves independently over each, and fuses the resulting ranked lists before reranking. Starting from OmniEmbed, the weakest first-stage baseline (0.17), this strategy closes the gap: MARQUIS-2 reaches 0.759. The tight clustering of MARQUIS variants (0.746–0.759), however, suggests that the gains come mainly from query expansion and reranking, not from the choice of fusion strategy.

Several questions remain. Combining a stronger first-stage model like Wholembed-v3 with C2F-RAG's cognitive reranking, or with a trained video-native reranker, would help separate the contributions of recall from reasoning. More broadly, the

Rank	System	nDCG@10	nDCG@20	nDCG@100	R@10	R@20	R@100
–	*OmniEmbed*	*0.166*	*0.186*	*0.245*	*0.096*	*0.158*	*0.297*
–	*OmniEmbed + RankVideo*	*0.542*	*0.534*	*0.546*	*0.423*	*0.462*	*0.494*
–	*Mixedbread*	*0.717*	*0.706*	*0.748*	*0.604*	*0.634*	*0.741*
6	MARQUIS-6	0.746	0.757	0.799	0.636	0.711	0.818
5	MARQUIS-8	0.747	0.758	0.800	0.636	0.711	0.818
4	MARQUIS-4	0.754	0.765	0.807	0.641	0.716	0.823
3	MARQUIS-14	0.757	0.768	0.810	0.650	0.725	0.832
2	MARQUIS-2	0.759	0.771	0.811	0.652	0.730	0.832
1	**C2F-RAG**	**0.848**	**0.851**	**0.853**	**0.773**	**0.832**	**0.837**

Table 1: MAGMaR 2026 — Retrieval track leaderboard (sorted by nDCG@10). Baselines shown in italics above the rule. Note that we received 17 submission and all of them beat a strong baseline based on the winner of last year's shared task. As many submitted systems were slight variants of each other, we are only showing the top 6 systems from the leaderboard.

results point to a gap in multimodal embedding models and rerankers for complex retrieval over long-form video.

4 Generation Task

Task. The generation task requires systems to produce a cited, persona-consistent article grounded in a set of relevant videos. In the *oracle* setting, systems receive the ground-truth relevant videos for each query, isolating generation quality from upstream retrieval variance. Generated articles are evaluated both by human annotators and by MiRAGE (Martin et al., 2026), an automatic framework that measures factuality, information coverage, groundedness, and citation attribution. Each MiRAGE entailment judgment is produced by CLUE (Zhang et al., 2026), a calibrated multimodal uncertain-inference model. Full results are in Table 2.

4.1 Systems

We provide Collaborative Article Generation (CAG) (Martin et al., 2025) as a baseline with Qwen3.5 as the backbone (Qwen Team, 2026). CAG achieves a human score of 3.088 with strong precision on both information (0.76) and citation (0.62) metrics, but low recall.

Four teams submitted 16 predictions to the leaderboard: TRACE (Yan et al., 2026), CRAFT (Bhosale et al., 2026), C2F-RAG (Dai et al., 2026), and MARQUIS (Chakraborty et al., 2026). TRACE builds a structured text timeline for each video via OCR and object detection, then uses a text-only LLM to localize relevant moments before passing selected frames to a vision-language model for claim generation. CRAFT generates atomic claims from query-conditioned keyframes and ASR, then iteratively verifies and repairs each claim through a hybrid critic loop before merging citations. C2F-RAG applies its retrieval pipeline's prompt sculpting mechanism to constrain generation to the persona and enforce chunk-level temporal grounding. MARQUIS explores multiple generation strategies over a shared evidence extraction layer, including direct QA, clustering-based summarization, and an RLM controller that iteratively gathers and curates evidence before writing.

4.2 Results

The generation results (Table 2) show several patterns across the submissions

Human preference vs. automatic metrics. The two top-ranked systems by human score, MARQUIS-iter-qa-base (3.833) and TRACE-04 (3.825), achieve relatively modest MiRAGE precision and recall compared to lower-ranked systems. Conversely, the CAG baseline and MARQUIS-Ginger achieve the highest InfoP (0.764 and 0.776) and CiteP (0.617 and 0.643), yet rank 6th and below by human score. The gap comes mainly from how QA-based systems handle unanswerable sub-questions: when the video evidence is insufficient, these systems refuse to answer rather than hallucinate, and human annotators reward this. MiRAGE, however, scores refusals as missing information and penalizes recall. So the systems human annotators prefer most, precisely because they acknowledge gaps rather than fabricate claims, are the ones automatic evaluation penalizes most.

QA-based generation. The MARQUIS QA variants dominate the top of the human leaderboard, holding three of the top four positions. These systems decompose the query into sub-questions, an-

Rank	System	Human	Best Votes	Best %	InfoP	InfoR	CiteP	CiteR
–	*CAG (baseline)*	*3.088*	*1*	*1.8%*	*0.764*	*0.410*	*0.617*	*0.228*
16	CRAFT-01	2.246	0	0.0%	0.695	0.389	0.065	0.000
15	CRAFT-02	2.281	0	0.0%	0.627	0.304	0.530	0.198
14	C2F-RAG-01	2.456	6	10.5%	0.557	0.466	0.452	0.349
13	C2F-RAG-02	2.526	2	3.5%	0.584	0.450	0.479	0.347
12	CRAFT-03	2.542	0	0.0%	0.609	0.304	0.509	0.204
11	TRACE-02	2.579	0	0.0%	0.621	0.550	0.536	0.500
10	TRACE-01	2.579	0	0.0%	0.599	0.564	0.515	0.498
9	MARQUIS-Bullet	2.667	0	0.0%	0.711	0.394	0.604	0.237
8	TRACE-03	2.702	2	3.5%	0.599	0.553	0.507	0.503
7	MARQUIS-ss-qa-base	3.070	6	10.5%	0.331	0.306	0.277	0.281
6	MARQUIS-Ginger	3.123	6	10.5%	0.776	0.404	0.643	0.226
5	MARQUIS-RLM	3.298	3	5.3%	0.708	0.385	0.592	0.272
4	MARQUIS-ss-qa-ginger	3.421	10	17.5%	0.544	0.324	0.326	0.238
3	MARQUIS-iter-qa-ginger	3.694	5	8.8%	0.345	0.290	0.257	0.226
2	TRACE-04	3.825	8	14.0%	0.640	0.483	0.498	0.405
1	MARQUIS-iter-qa-base	3.833	8	14.0%	0.347	0.313	0.268	0.258

Table 2: MAGMaR 2026 — Oracle generation track leaderboard (sorted by human score). Baseline shown in italics above the rule. 4 teams submitted systems. All teams had at least one system which earned best votes.

swer each from the video evidence, and synthesize the results into an article. Annotators find these outputs more complete and easier to evaluate, likely because the question-answer structure gives the article a natural organization. The gap between iterative and single-shot QA variants (3.833 vs. 3.070) suggests that iterating over the evidence, asking follow-up questions and refining answers, produces noticeably better articles.

Evidence verification. Systems with explicit claim verification show stronger citation precision. CRAFT-03 achieves a CiteP of 0.509 despite ranking 12th overall, and MARQUIS-Ginger reaches 0.643 with its CLUE-based calibration filtering. The TRACE submissions achieve the most balanced precision-recall profiles across both information and citation metrics: TRACE-04 reaches 0.640 InfoP / 0.483 InfoR and 0.498 CiteP / 0.405 CiteR, the strongest combined recall among top-ranked systems.

The shared task results point to several open problems. The clearest is the disconnect between human preference and automatic metrics: systems that abstain when evidence is insufficient are rewarded by annotators but penalized by MiRAGE, which treats all omissions as precision failures. Evaluation frameworks that account for appropriate refusal and persona-relevant (not merely factual) coverage are a clear next step. On the modeling side, no submitted system uses task-specific fine-tuning; the strong performance of QA-based approaches suggests that models trained for grounded multi-video question answering with persona conditioning could improve results. Finally, all current systems convert videos to text proxies before reasoning, discarding visual information that captions cannot capture. Architectures for direct multi-video visual reasoning are needed to close this gap.

5 Conclusion

The shared task at the second workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2026) attracted two retrieval teams and four generation teams. Text-based reasoning dominates both tracks: C2F-RAG achieves the highest retrieval scores by indexing summaries and captions rather than operating on video directly, and every generation system converts videos to text before reasoning. Given this shared substrate, inference-time reasoning proves more important than the choice of first-stage model. Reranking transforms OmniEmbed from the weakest retrieval baseline into a competitive first stage, and iterative QA generation produces noticeably better articles than single-shot variants. On the generation side, human preference and automatic metrics also diverge: systems that appropriately abstain when evidence is insufficient are rewarded by annotators but penalized by MiRAGE, which treats all omissions as failures.

These results point to several directions for fu-

ture work. Evaluation frameworks need to account for appropriate refusal and persona-relevant coverage beyond factual completeness. No submitted system uses task-specific fine-tuning in either track, and the strong performance of QA-based generation suggests that models trained for grounded, persona-conditioned multi-video question answering could improve further. All current systems also discard visual information by converting videos to text before reasoning; architectures for direct multi-video visual reasoning remain unexplored.

References

Mahesh Bhosale, Abdul Wasi, Vishvesh Trivedi, Pengyu Yan, Akhil Gorugantu, and David Doermann. 2026. Craft: Critic-refined adaptive key-frame targeting for multimodal video question answering. *Preprint*, arXiv:2605.19075.

Debashish Chakraborty, Dengjia Zhang, Jialiang Jin, Hanting Liu, Katherine Guerrerio, Hanxiang Qin, Tyler Skow, Alexander Martin, Reno Kriz, and Benjamin Van Durme. 2026. Marquis: A three-stage pipeline for video retrieval-augmented generation. *Preprint*, arXiv:2605.17640.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2025. M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Jiaxin Dai, Zehang Wei, Jiamin Yan, and Xiang Xiang. 2026. Decoupling semantics and logic: A training-free coarse-to-fine pipeline for video retrieval-augmented generation. *Arxiv preprint*, 2606.07924.

Reno Kriz, Kate Sanders, David Etter, Kenton Murray, Cameron Carpenter, Kelly Van Ochten, Hannah Recknor, Jimena Guallar-Blasco, Alexander Martin, Ronald Colaianni, Nolan King, Eugene Yang, and Benjamin Van Durme. 2025. Multivent 2.0: A massive multilingual benchmark for event-centric video retrieval. *Preprint*, arXiv:2410.11619.

Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, Luca Soldaini, and Eugene Yang. 2025. Overview of the trec 2024 neuclir track. *Preprint*, arXiv:2509.14355.

Xueguang Ma, Luyu Gao, Shengyao Zhuang, Jiaqi Samantha Zhan, Jamie Callan, and Jimmy Lin. 2025. Tevatron 2.0: Unified document retrieval toolkit across scale, language, and modality. *Preprint*, arXiv:2505.02466.

Sean MacAvaney, Craig Macdonald, and Iadh Ounis. 2022. Streamlining evaluation with ir-measures. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II*, volume 13186 of *Lecture Notes in Computer Science*, pages 305–310. Springer.

Alexander Martin, Reno Kriz, William Gantt Walden, Kate Sanders, Hannah Recknor, Eugene Yang, Francis Ferraro, and Benjamin Van Durme. 2025. Wikivideo: Article generation from multiple videos. *Preprint*, arXiv:2504.00939.

Alexander Martin, William Walden, Reno Kriz, Dengjia Zhang, Kate Sanders, Eugene Yang, Chihsheng Jin, and Benjamin Van Durme. 2026. Seeing through the mirage: Evaluating multimodal retrieval augmented generation. *Preprint*, arXiv:2510.24870.

Mixedbread Team. 2026. Beyond the limit: Introduce mixedbread wholembed v3. Blog post.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Tyler Skow, Alexander Martin, Benjamin Van Durme, Rama Chellappa, and Reno Kriz. 2026. Rankvideo: Reasoning reranking for text-to-video retrieval. *Preprint*, arXiv:2602.02444.

Pengyu Yan, Akhil Gorugantu, Mahesh Bhosale, Abdul Wasi, Vishvesh Trivedi, and David Doermann. 2026. Trace: Evidence grounding-guided multi-video event understanding and claim generation. *Preprint*, arXiv:2605.16740.

Dengjia Zhang, Alexander Martin, William Jurayj, Kenton Murray, Benjamin Van Durme, and Reno Kriz. 2026. Unified multimodal uncertain inference. *Preprint*, arXiv:2604.08701.

A WIKIVIDEO25 Dataset

WIKIVIDEO25 builds on WIKIVIDEO (Martin et al., 2025), which contains 57 English topics from 2015–2023, with ∼7 videos per topic. It has several limitations: it is English-only, its articles are written top-down from preexisting Wikipedia lead sections, its videos are mostly professional news broadcasts, and its queries are underspecified (e.g., "tell me about [event name]") and do not reflect realistic information-seeking behavior (Lawrie et al., 2025). Because the topics are well-known events, the relevant information often already exists in pretrained model weights. In some cases, simply quoting the Wikipedia lead section outperforms human-written responses. WIKIVIDEO25 targets each of these problems.

Queries and Personas. Each topic comes with a persona and a query (two per topic; see Figure 2). The persona is a detailed background for the query provider. It constrains what counts as relevant evidence, so systems must reason about which parts of

the videos matter for a specific professional need, not just surface everything about the event. Each query is a complex information need written in one of two ways: *biased*, where the author had watched the videos and knows what the collection contains, and *unbiased*, where the author had only seen video titles and writes a more natural request without that knowledge. Unbiased queries may ask for information absent from the collection, so they also test whether systems can acknowledge gaps in the evidence instead of hallucinating an answer.

Video-Centric Reports. WIKIVIDEO's articles were written top-down: annotators began with the Wikipedia lead section for each event and removed claims not supported by the videos. The ground-truth articles were therefore anchored to an external text source, not to the video content itself, and in some cases quoting the Wikipedia lead outperformed human-written responses. WIKIVIDEO25 reverses this. Annotators first write atomic claims grounded in each video's audiovisual content, then combine them into higher-level inferences that form the final article. This bottom-up annotation produces reports more faithful to the video content and reduces the advantage of parametric knowledge, since the ground truth is no longer derivable from a well-known Wikipedia article.

Knowledge cutoff. All events in WIKIVIDEO25 take place in 2025, after the training cutoffs of the models used in submitted systems. Because none of the event-level facts exist in these models' parametric knowledge, systems cannot shortcut the task by recalling memorized information and must ground their outputs in the retrieved video evidence. This is resolves the quoting issue noted above from WIKIVIDEO, whose topics were well-known events with extensive coverage in pretraining data. Using events beyond any model's knowledge cutoff removes this issue. It also means the MultiVENT 2.0 distractor set, which covers events only through 2023, contains no true positives for any WIKIVIDEO25 query. The result is a cleaner evaluation of both retrieval and generation: retrieval systems must identify relevant videos rather than match familiar event descriptions, and generation systems must synthesize articles from audiovisual content rather than parametric knowledge.

Raw Videos. The original dataset is mostly professional news broadcasts, where a third-party narrator describes the event to the viewer. For these, an audio transcript alone is often enough to produce a reasonable article. WIKIVIDEO25 shifts toward raw and "diet raw" footage: videos captured by participants or bystanders without professional narration or post-production editing. These videos are more ambiguous and force models to infer what is happening from low-level audiovisual cues. There is no narrator explaining the scene.

Multilingual Videos. WIKIVIDEO25 adds two topics with Mandarin-language videos and one mixed Nepali–English topic. Systems must now process non-English audio and OCR at both retrieval and generation. In the mixed-language case, they must also reason across videos in different languages about the same event.

Author Index