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,

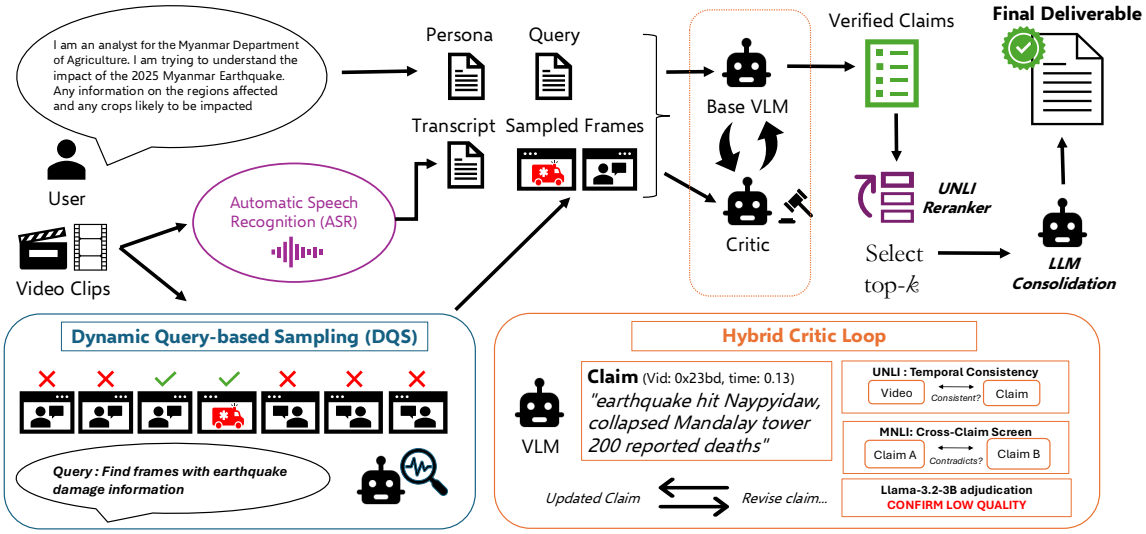

Figure 1: **Overview of CRAFT.** Given a persona, query, and relevant videos, CRAFT builds a query-specific multimodal evidence stream: each video is transcribed once via ASR, and *Dynamic Keyframe Selection* (DKS) selects the frames most relevant to the query. The base VLM consumes the persona, query, 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.