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

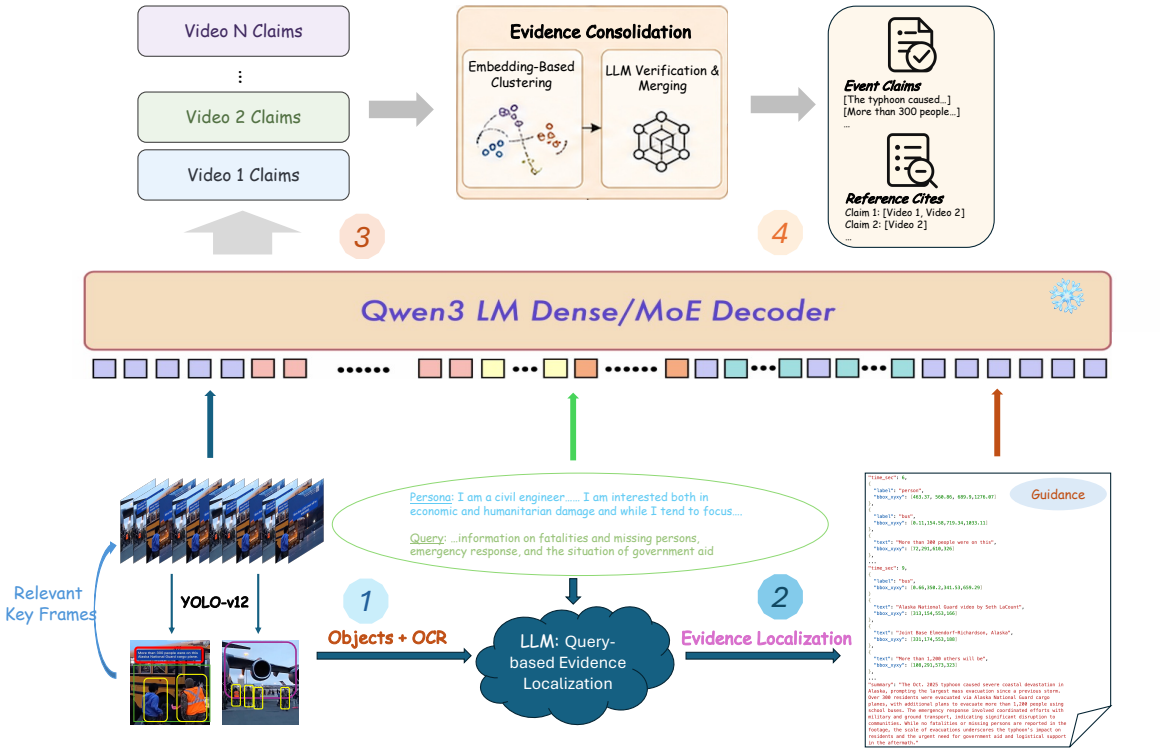

Figure 1: **Grounding-guided pipeline for event video claim generation.** We extract structured grounding signals via object detection and OCR over video frames, then use a text-only LLM to align detected labels and on-screen text with the query and persona to identify relevant moments. This text-based 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.