WikiVideo: Article Generation from Multiple Videos

Anonymous ACL submission

Abstract

We introduce the task of grounded article generation with the goal of creating a Wikipedia-style article from multiple diverse videos about real-world events—from natural disasters to political elections—where all the information in the article is supported by video evidence. Videos are intuitive sources for retrieval-augmented generation (RAG), but most contemporary RAG workflows focus heavily on text while existing methods for video-based summarization focus on low-level scene understanding rather than high-level event semantics. To close this gap, we introduce WIKIVIDEO, a benchmark consisting of expert-written articles and densely annotated videos that provide evidence for articles' claims, facilitating the integration of video into RAG pipelines and enabling the creation of in-depth content that is grounded in multimodal sources. We further propose **C**ollaborative **A**rticle **G**eneration (CAG), a novel interactive method for article creation from multiple videos. CAG leverages an iterative interaction between an r1-style reasoning model and a VideoLLM to draw higher-level inferences about the target event than is possible with VideoLLMs alone, which fixate on low-level visual features. We benchmark state-of-the-art VideoLLMs and CAG in both oracle retrieval and RAG settings and find that CAG consistently outperforms alternative methods, while suggesting intriguing avenues for future work.

1 Introduction

Audiovisual media is becoming an increasingly dominant form of online information consumption. From firsthand, "in the wild" video footage of natural disasters to professionally edited news coverage of major political events, videos serve as rich sources of information for producing factual, grounded articles. Especially for *actively unfolding* events, grounding articles in video not only can potentially combat misinformation among readers, but can also provide a useful tool for journalists and other writers to quickly synthesize information about new developments. Figure 1 motivates this task, taking an information request and producing a Wikipedia-style article with references that ground the article in the supporting video content.

Current methods and resources for article generation overwhelmingly rely on textual sources (Liu et al., 2018a; Barham et al., 2023; Lawrie et al., 2024; Shao et al., 2024, *i.a.*), while video understanding benchmarks largely focus on low-level tasks such as entity-centric question answering or captioning (Xu et al., 2016; Krishna et al., 2017; Yu et al., 2019; Li et al., 2022; Lin et al., 2024). Kriz et al. (2024) show that models trained on such tasks often fail in more realistic settings that require understanding of high-level semantics, (e.g.) as conveyed in articles about actual events. The MultiVENT benchmarks (Sanders et al., 2023; Kriz et al., 2024) are distinctive in focusing on *major real-world events* as depicted in *multiple* videos—spanning firsthand footage, amateur-edited clips, and news broadcasts.

In this work, we introduce WIKIVIDEO, a benchmark that builds upon MultiVENT and evaluates the ability to write event-centric articles in the style of a Wikipedia lead (overview) section based *only* on video content. Given a request about a real-world event, systems must retrieve a set of relevant videos and then generate an article from the videos' (visual, audio, and OCR) content, providing grounding references (citations) to where the information comes from. WIKIVIDEO consists of 57 events and 427 relevant videos (from a corpus of 109K; avg. 8 relevant/event), with expert-written reference articles that synthesize information about each event across *all* relevant videos—forcing systems to not only understand high-level information within a *single* video, but also to synthesize information *across* multiple videos on the same topic. Systems able to achieve strong results on

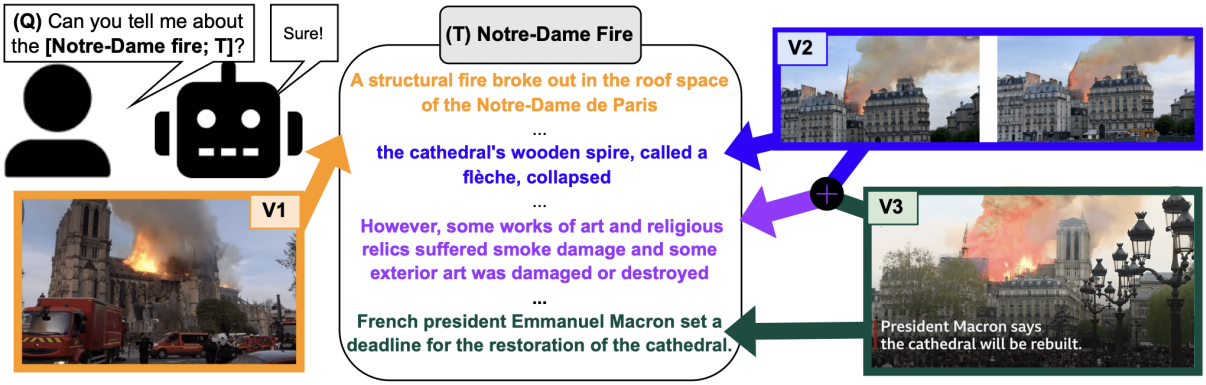

Figure 1: WIKIVIDEO introduces the task of Article Generation from Multiple Videos, which requires writing a high-level article in the style of a Wikipedia lead, given a target event (T), a query about that event (Q), and a collection of Q-relevant videos (V). All claims in the article are grounded in visual, audio, and/or OCR content of video(s) in V (indicated by matching colors between text and frame borders above).

WIKIVIDEO would be of significant practical use in grounded article generation from multimodal sources, and would enable both the rapid seeding of new Wikipedia articles for actively unfolding events and the enriching of existing articles with audiovisual content.

To support the WIKIVIDEO task, we propose CAG (**C**ollaborative **A**rticle **G**eneration), a novel method capable of generating high-level articles. Inspired by relevance feedback (Rocchio, 1971) and recent advances in *test-time scaling* (DeepSeek-AI et al., 2025), CAG involves collaborative interaction between (1) a VideoLLM, (2) a text-based reasoning model, and (3) a text-only LLM to extract information from videos and to aggregate the information into an article. The VideoLLM extracts low-level information, such as on-screen text and descriptions of visual entities, while the reasoning model provides relevance feedback on the extracted information, optionally requesting new extractions from the VideoLLM before feeding *relevant* ones to the text-only LLM. The LLM then aggregates the relevant extractions, drawing higher-level inferences about the underlying event, in order to generate the final article. We summarize our contributions as follows:

1. We introduce WIKIVIDEO, a new dataset and task for generating articles from multiple videos. WIKIVIDEO is the first benchmark for multi-video article generation, requiring reasoning across audio and visual information, covering 57 events (topics) as depicted in 427 videos that are densely annotated with modality-specific claim grounding annotations, with expert-written reference articles for each event.

2. We introduce CAG, a novel method for article generation from multiple videos that is based on relevance feedback and test-time scaling.

3. We present a broad suite of experiments that evaluate both CAG and popular VideoLLMs on WIKIVIDEO across a range of settings, demonstrating CAG's superiority to other approaches while revealing WIKIVIDEO to be a challenging benchmark.

2 Related Work

Video Understanding and Summarization Video summarization has been studied on small-scale video datasets, such as SumMe (25 videos; Gygli et al., 2014) and VideoSum (50 videos; Song et al., 2015)—considerably smaller than WIKIVIDEO ($\sim$ 400 videos). Work on *cross-modal* summarization that leverages video largely focuses on producing low-level scene descriptions as the summary, since the associated tasks are chiefly concerned with *aligning* video scenes with caption-like text (He et al., 2023; Lin et al., 2024; Hua et al., 2024). Other work treats video summaries as mere LLM syntheses of frame-level captions (Hua et al., 2024; Zhang et al., 2024a). In contrast, WIKIVIDEO is focused on summaries that provide *high-level* information *supported by* video content.

Ren et al. (2025) recently proposed the task of video retrieval augmented generation (videoRAG) over long videos. While we too explore retrieval in our experiments, our data and task are considerably different: whereas Ren et al. exclusively use highly polished videos (e.g. documentaries, lectures) and only single videos, much of WIKIVIDEO consists of raw and amateur-edited footage of events *in the*

wild and *in real time* and invovles reasoning across information sources. Further, whereas they are concerned with short-form question answering, we are concerned with long-form article generation.

Similarly, there is much other video understanding work oriented toward tasks other than summarization, such as retrieval of (Chen and Dolan, 2011; Xu et al., 2016; Anne Hendricks et al., 2017; Wang et al., 2020) question answering about (Jang et al., 2017; Lei et al., 2018; Yu et al., 2019), and recognition of (Zhou et al., 2019; Sanders et al., 2024) low-level video features and concepts that span a few seconds or exist only at the frame level.

Article Generation has largely been studied in text-only settings and as a multi-document summarization task. Early work in this area was performed as part of the DUC and TAC conferences[1], including the DUC 2003-2007 multi-document summarization tasks and the TAC 2010 and 2011 Guided Summarization track. Whereas this early work—and many more recent efforts (Hermann et al., 2015; Nallapati et al., 2016; Fabbri et al., 2019; Huang et al., 2024, *i.a.*)—focused on summarizing *news* articles, generation of Wikipedia-style articles has received increasing attention. To our knowledge, (Sauper and Barzilay, 2009) were the first to have attempted this, focusing on generation of full Wikipedia articles by filling learned article templates with sentences extracted from source articles. Similar to us, (Liu et al., 2018b) focus on Wikipedia lead sections, taking article titles and a collection of source documents as input and performing abstractive summarization using a decoder-only Transformer and introducing the WikiSum dataset as part of their work. (Zhu et al., 2021) also focus on leads, but take a topic modeling-inspired approach, assigning topics to source article paragraphs and conditioning generation of each lead sentence on paragraphs associated with either a single predicted topic or a mixture of topics. Like Sauper and Barzilay (2009), Shao et al. (2024) tackle full Wikipedia article generation, using a complex pipeline that entails (1) surveying related Wikipedia articles, (2) generating perspectival questions and answers via simulated dialogues, (3) using these dialogues to construct an outline for the article, and (4) populating the outline from section titles and headings of source articles retrieved during (2). Concurrently, Yang et al. (2025) explore multimodal article generation, but aimed at incorporating figures into articles rather than synthesizing information.

Finally, in focusing on Wikipedia articles about *events*, we extend a recent line of work on explicitly *event-centric* summarization, in which generations must cover relevant information about a *single* target event (Vallurupalli et al., 2022; Gantt et al., 2024; Walden et al., 2024).

3 WIKIVIDEO Dataset

WIKIVIDEO is built using videos from MultiVENT 1.0 (Sanders et al., 2023) and MultiVENT 2.0 (Kriz et al., 2024) that are linked to Wikipedia articles obtained from the May 2025 dump provided by the MegaWika2 dataset (Barham et al., 2025). Our data collection process consists of five steps: (1) initial event and article selection, (2) article claim decomposition, (3) claim correction, (4) claim grounding, and (5) article rewriting. Figure 2 illustrates the core components of the annotation process: decomposition, grounding, and rewriting.

Initial Event and Article Selection We select an initial set of events from MultiVENT subject to two constraints: the event must (1) have English-language videos associated with it and (2) must have a link to an English Wikipedia article about the event.[2] In total, there are 63 events in MultiVENT satisfying both criteria, supported by 503 associated videos.

For each, we extract the lead section from the linked Wikipedia article to use as the basis for the remainder of our annotation. Lead sections are distinctly suited to our goal of curating *high-level* articles, as they are intended to provide a summary of the most important aspects of the entire page.[3]

Claim Decomposition and Correction Next, following recent work on *claim decomposition* (Min et al., 2023; Wanner et al., 2024b; Gunjal and Durrett, 2024, *i.a.*), we decompose each sentence of the Wikipedia lead section into a set of contextualized, atomic *subclaims* via few-shot prompting of Qwen 2.5 32B (Qwen et al., 2025). Expert annotators versed in the claim decomposition literature then manually correct these decompositions to ensure atomicity and faithfulness to the original text.[4]

[1] DUC: https://duc.nist.gov/; TAC: https://tac.nist.gov

[2] As MultiVENT is *multilingual*, not all events it contains satisfy (1) and many also do not satisfy (2).

[3] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section

[4] Appendix B for details on subclaim decomposition/correction.

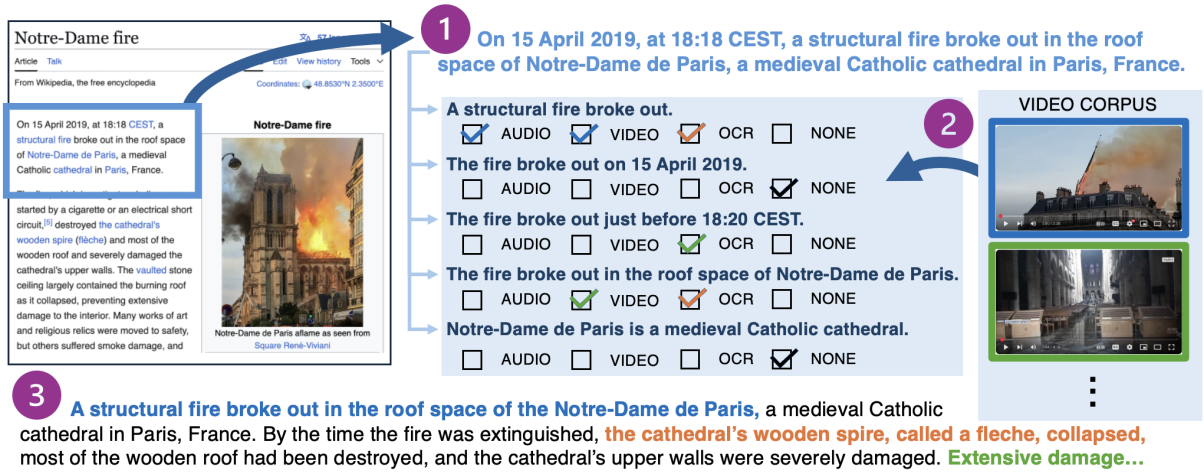

Figure 2: The WIKIVIDEO curation process. (1) Sentences in Wikipedia lead sections are decomposed into subclaims. (2) Subclaims are grounded in audio, video, and/or OCR evidence. (3) Leads are rewritten to cover only the *grounded* information.

WIKIVIDEO	
Video Length (s)	79.6
Article Length (toks)	118
Videos	7.65
Audio Subclaims	25.0
Video Subclaims	18.3
OCR Subclaims	28.5
A/V/O Subclaims	13.0
Total Subclaims	51.1

Table 1: Dataset (top) and per-event (bottom) averages.

Subclaim Grounding Given the corrected subclaims for the Wikipedia lead section associated with each event, we next attempt to *ground* the subclaims in relevant videos. This grounding task provides *modality-specific* annotations for each subclaim and for each video associated with the target event, as annotators were asked to indicate whether a subclaim is supported by the video's (non-text) visual content, its OCR content, its audio content, or none of the above. In-domain expert annotators completed these annotations for all 58 events and 503 videos. Our pilot with these annotators obtained a high overall agreement ($\alpha = .767$).

Article Rewriting Finally, given the grounded set of subclaims and their corresponding videos, three of the authors rewrote the Wikipedia lead sections such that the resulting articles contained all and only information supported by the video-grounded subclaims. During this stage, six events were found to have too few grounded subclaims to support a rewritten article of any substance, and were subsequently removed from the final dataset.

Final Dataset The final WIKIVIDEO dataset consists of 57 events (topics) spanning 427 videos annotated with grounded subclaims, where each event is associated with a fully grounded, expert-written article. The events in WIKIVIDEO span from 2016-2025, with a specific subset being outside the parametric memory of current (V)LMs. Table 1 provides a summary, with more in Appendix A & B.

4 Article Generation from Multiple Videos

Task The WIKIVIDEO article generation task takes as input (1) a topic event T, (2) a query about T, and (3) a set of videos $V = \{v_1, \ldots, v_n\}$ deemed relevant to (i.e. depicting some facet of) T. The output is then a natural language article A_p generated conditional on T, Q, and V. A_p must be fully grounded in V, providing citations to videos that support the claims in A_p. In this work, we consider two possible sources for V: the reference set of videos for T as annotated in MultiVENT 1.0 and 2.0 (the *oracle* setting) and a set of videos obtained from a retrieval model (the *RAG* setting).

4.1 Collaborative Article Generation (CAG)

Overview Conditional text generation from multiple videos faces several challenges that hinder the efficiency and effectiveness of current methods. First, open-source VideoLLMs are generally trained to produce low-level scene descriptions, making extraction of high-level concepts (necessary for complex event understanding) a challenge, even based on a single video—let alone

multiple. Second, running inference over multiple long videos is memory-intensive. For instance, in preliminary experiments with several of the VideoLLMs we consider in even 8 80GB A100s struggled to accommodate a single long video (5+ minutes) at 1 fps, as well as two or more videos at 0.25 fps, additionally supported by the findings of (Li et al., 2025).

To help address these limitations, we introduce **C**ollaborative **A**rticle **G**eneration (CAG; Figure 3), a method for article generation from multiple videos that draws on recent developments in *test-time scaling* (DeepSeek-AI et al., 2025; Huang et al., 2025; Weller et al., 2025; Jurayj et al., 2025) in addition to the classic notion of *relevance feedback* from information retrieval (Rocchio, 1971). CAG features three core components: a VideoLLM, a reasoning model, and an LLM.

Collaborative Per-Video Summarization The first phase of CAG involves a collaborative, iterative exchange between the VideoLLM and the reasoning model. The VideoLLM begins by generating generic summaries of each video based on a simple prompt to "describe the video in detail." The resulting summaries provide salient low-level information, covering scene descriptions and prominent on-screen text.

Next, the reasoning model assists the VideoLLM in producing a *refined* summary for each video that covers higher-level information about the underlying event. Concretely, the reasoning model is given both Q (here, the name of the target event T) and the initial generic summary for a particular video as input, and is then asked to either: (1) return the original summary if the reasoning model deems it to be adequate, or else (2) generate a *new* prompt seeking additional information about T not attested in the input summary. This new prompt is then used to elicit a refined summary from the VideoLLM—an action we dub REPROMPTING. The reasoning model can thus be understood as providing a form of relevance feedback on the VideoLLM-generated summary with respect to Q (and thus to T). This process is iterative because the reasoning model may in principle REPROMPT repeatedly—requesting new summaries from the VideoLLM that are more relevant to Q until it is satisfied with the result. In practice, we enforce a maximum REPROMPTING *iteration budget*—analogous to *test-time compute* budgets for recent reasoning models—that a higher budget tends to yield higher-quality articles.

Article Synthesis Once the reasoning model determines that the current query is adequate (or the REPROMPTING iteration budget is exhausted), the final article is synthesized using a text-only LLM. This model takes as input (1) the original generic summary output by the VideoLLM for each video; (2) all REPROMPTED VideoLLM queries and their resulting (more event-targeted) summaries; and (3; optionally) an audio transcript of each video. Given these inputs, the model is then instructed to generate the article about T.

We note that in virtue of REPROMPTING—enabling a reasoning model to iteratively craft prompts for the VideoLLM in order to produce summaries more explicitly targeted to an event of interest (T)—CAG goes some way toward mitigating the problem of summaries that are overly focused on low-level descriptions. Further, in processing one video at a time, CAG reduces the memory burden of simultaneous multi-video inference.

5 Experiments

We conduct experiments on WIKIVIDEO that (1) benchmark CAG against baseline VLM approaches; (2) demonstrate the strengths of CAG over LLM approaches in generalization to unseen events and grounding information in videos; (3) asses the impact of including raw audio and audio transcripts as input to the text-only LLM; and (4) evaluate the effectiveness of different retrievers in the RAG setting, where videos must first be retrieved. (1-3) are conducted in the *oracle* setting, where V consists of all and only relevant videos.

Models For VideoLLMs, we consider LLaVA-Video-72B ("LLaVA-Video" or "LV" in results; Zhang et al., 2024b), VAST (Chen et al., 2024), InternVideo2.5-8B (InternVideo2.5 or IV; Wang et al., 2025), and QwenVL2.5-72B (Qwen2.5VL or QVL; Bai et al., 2025). We use DeepSeek-R1 distilled to Qwen-32B (DeepSeek-AI et al., 2025) as the reasoning model and Qwen2.5 (Qwen et al., 2025) as the text-only LLM.

Metrics We use a suite of different metrics to evaluate the generated articles. We first present ROUGE-{1,2,LCS} F_1 (R1, R2, RL; Lin, 2004) and BERTScore F_1 (BS; Zhang et al., 2019) as two widely used metrics for free-form text generation, using the human-written articles as references. As these metrics largely focus on *lexical* similarity, we

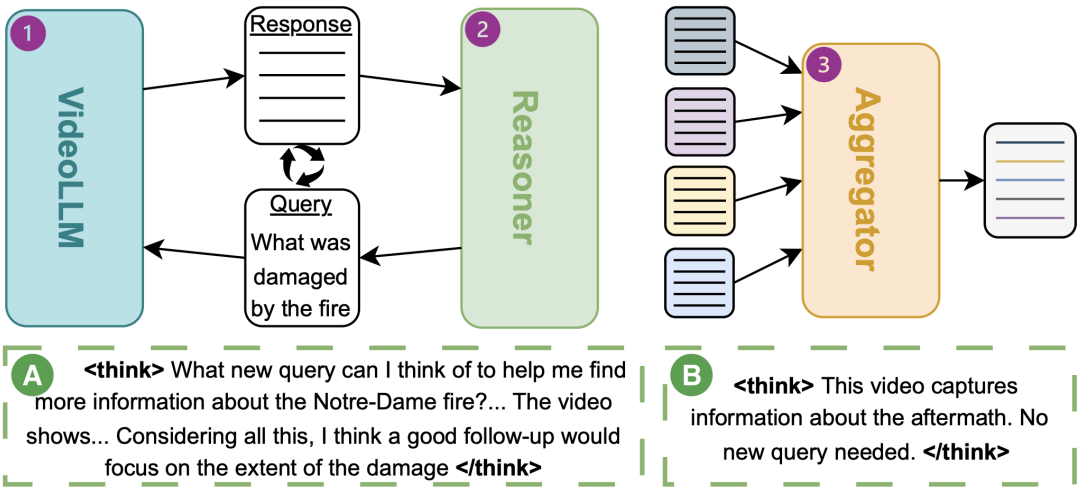

Figure 3: CAG involves an iterative exchange between (1) a VideoLLM that generates per-video summaries and (2) a reasoning model that evaluates them and produces more event-targeted prompts that are then fed back to the VideoLLM to obtain more comprehensive summaries. Finally, a text-only LLM (3) aggregates these summaries into an full article. Boxes A and B show shortened reasoning chains from the reasoner.

Method	VideoLLM	R1	R2	RL	BS	Arg	AS
CONCAT-0	LLaVA-Video	7.34	1.60	4.78	71.99	19.31	5.08
	InternVideo2.5	11.85	2.32	7.90	80.78	18.33	9.53
	QwenVL2.5	11.34	3.13	7.06	81.60	23.72	8.01
CONCAT-R	LLaVA-Video	6.36	1.51	4.22	80.03	21.34	5.50
	InternVideo2.5	6.93	1.68	4.83	79.62	22.48	6.19
	QwenVL2.5	8.38	2.71	5.49	81.94	22.89	7.17
CAG-0	LLaVA-Video	30.02	8.68	17.59	77.59	26.21	13.51
	InternVideo2.5	32.54	8.98	19.47	85.82	25.65	**17.58**
	QwenVL2.5	33.58	10.15	19.15	86.18	28.97	15.63
CAG-R	LLaVA-Video	33.38	10.05	19.44	84.55	28.26	15.23
	InternVideo2.5	33.91	9.58	**20.07**	86.13	27.01	14.23
	QwenVL2.5	**33.96**	**10.90**	19.45	**86.35**	**30.77**	14.29

Table 2: WIKIVIDEO article generation results for CAG and baselines *without* audio inputs (i.e. vision only). CAG obtains the best results (**bolded**) across metrics, with performance increasing with # iterations (2) for most metrics.

additionally present AlignScore (AS; Zha et al., 2023), a metric for factual consistency based on a learned text-pair alignment function that outputs a scalar value in $[0, 1]$ representing the degree of information alignment between the two texts.

Lastly, we also evaluate the extent to which predicted articles recover *specific* pieces of event-relevant information. We map each WIKIVIDEO event into the seven-type event ontology defined in the MultiVENT-G dataset (Sanders et al., 2024), each of which is associated with a set of role-focused questions about events of that type.[5] We use an LLM (GPT-4o) to extract answers to these questions from both the reference and predicted articles. Since a question may have multiple answers, we compute a maximum bipartite matching between predicted and reference answers, obtaining an alignment between them that optimizes normalized edit distance between paired answer spans. We then report an answer span F_1 given this alignment, using normalized edit distance in lieu of (overly stringent) exact match. Prior work on event extraction has leveraged similar metrics to evaluate event *argument* F_1 (Du et al., 2021; Chen et al., 2023a,b; Vashishtha et al., 2024), so we refer to this as "Arg."

[5]The event types are Sporting Events, Natural Disasters, Elections, Social Events, Demonstrations, Discoveries/Launches, and Political Developments. Event specific scores in Appendix G.

Baselines We consider several baseline article generation methods that ablate different components of CAG. The first baseline (CONCAT-0) simply concatenates the generic per-video summaries to produce the final article, ablating both the aggregator and reprompting. The second (CONCAT-REPROMPT) concatenates only the per-video REPROMPTED summaries, excluding the generic ones, while still ablating the aggregator. The third, (CAG-0), uses the aggregator but fixes CAG's iteration budget to 0—relying exclusively on the generic per-video summaries. The comparison between CAG-0 and CAG-R (iteration budget of 2) thus offers an illustration of test-time scaling of CAG via a larger iteration budget.

5.1 CAG and Baselines

Table 2 shows WIKIVIDEO article generation results comparing CAG-R against the baselines described above. We find that simple concatenations of the per-video summaries—whether the initial generic ones (CONCAT-0) or those obtained via reprompting (CONCAT-REPROMPT, CONCAT-R)—yield articles of poor quality. Although manual inspection reveals these per-video summaries to contain mostly accurate descriptions of scenes and notable visual entities (e.g. the Eiffel tower), we take this as compelling evidence that individual video summaries are inadequate for our task, absent higher-level synthesis.

Results with CAG-0 and CAG-R, both of which incorporate the text-only aggregator LLM, offer further evidence for this interpretation, as we observe large gains across all metrics for both of these methods relative to the CONCAT baselines. For most metrics, CAG-R also obtains superior results to CAG-0, suggesting that supplying the aggregator with the reprompted summaries (in addition to the generic ones) further enhances article quality.

5.2 Generalization to Unseen Events

LLMs are trained on vast quantities of internet text, including Wikipedia (Barham et al., 2023; Soldaini et al., 2024; Cheng et al., 2024). As such, the LLMs underlying each VLM in our experiments can be assumed to have parametric knowledge about WikiVideo events from 2016 to 2024. To illustrate the generalizability of CAG beyond these events, Table 3 reports results on a subset of WikiVideo containing events that occurred *after* the latest knowledge cutoff date of any of these models.[6]. Here, we find that CAG obtains superior results to the Qwen 2.5 LLM, indicating that it exhibits better generalization to genuinely novel events.

Method	R1	BS	Arg	AS	G
QVL	15.6	82.6	25.6	15.9	–
CAG	**42.0**	**85.7**	**32.8**	14.0	**51.2**
LLM	33.5	84.5	22.9	**18.4**	19.5

Table 3: Results on WIKIVIDEO-25 (only from 2025). LLM- Qwen 2.5 LLM. G: human annotated grounding.

This experiment also highlights an interesting finding: *VLMs struggle to connect visual signals to parametric knowledge.* This can be see in Table 3, where the performance of CONCAT-0 is similar for all events and the 2025 subset. Appendix E shows summaries for a single video about the 2019 Notre Dame Fire generated by QwenVL2.5 (left) and CAG (right). While the single video summary correctly identifies the location (Paris), the most salient entity (Notre Dame), and the physical event (fire), the output doesn't make any high-level inference about the event and it fails to connect the visual information to the event in parametric knowledge. We suggest two possible reasons for this: (1) the backbone LLM loses the ability to perform high-level inference because text-video pretraining data is purely focused on low-level descriptions (Zhang et al., 2024a), or (2) the LLM does not have the ability to activate this parametric knowledge from the projected visual signals.

5.3 Grounding in Videos

Information presented in the generated articles should be fully grounded in the videos provided as context. To evaluate this, we have human annotators perform the WikiVideo claim grounding annotation task on the model predictions from Table 3 for CAG and the Qwen2.5 LLM. We score the *G*roundedness of an article as the mean number of claims judged as supported by the video content: $G = \frac{1}{|C|}\sum_{c_i \in C} f(c_i, V)$, where C is the set of claims in the output, V is the set of associated videos, and f is scored as:

$$f(c, V) = \begin{cases} 1 & c \text{ is supported by some } v_i \in V \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

[6]Qwen2.5 Release: 2024, Earliest 2025 Event: January 15

Method	VLM	R1	BS	Arg	AS
CAT-0	QVL	11.1	81.6	22.1	7.50
CAG-R	QVL	32.1	85.7	26.3	12.6
CAG-R	QVL	**34.0**	**86.4**	**30.8**	**14.3**

Table 4: WIKIVIDEO article generation results for CAG *with* audio inputs. Bottom row shows CAG *without* audio (copied from Table 2), which performs best.

Retriever	VLM	R1	BS	Arg	AS
VC	QVL	24.1	83.7	20.9	10.6
MRF	QVL	23.8	77.9	20.7	9.0
Oracle	QVL	**34.0**	**86.4**	**30.8**	**14.3**

Table 5: Results with CAG using different retrievers. The top 5 videos in a ranked list are used for generation. Oracle retrieval results are from Table 2. VC: Video-ColBERT, MRF: MMMORRF

In Table 3, we find that a majority of the claims from CAG are grounded in the article—a stronger performance than the LLM-only generated responses. However, the difference between Rouge and BertScore does not reflect the poor quality of LLM generations or the lack of grounding, highlighting a need for strong evaluation metrics for video-to-text tasks.[7]

5.4 Experiment 5: Including Audio

Articles in WIKIVIDEO have many claims grounded partly or only in videos' audio signal (Table 1). Here, we consider the impact of adding audio information as additional input to CAG. In Table 16 we observe consistently worse results with audio inputs than without. We found that including audio transcripts tended to result in substantially shorter final articles (avg. $\sim$ 164 tokens vs. $\sim$ 206 tokens)—suggesting that they may be less thorough in their coverage of the event relative to the references. Identifying ways to more effectively incorporate audio into the WIKIVIDEO task is thus an intriguing direction for future work.[8].

5.5 Experiment 5: Retrieval Augmented Generation

In contrast to the previous two experiments, which were run using only relevant videos for each target event (the *oracle* setting), here we consider the *RAG* setting, in which relevant videos must be retrieved. We use the full set of MultiVENT 2.0 (Kriz et al., 2024) videos from the test set as our corpus (109K videos). We perform retrieval from this collection of videos with two retrievers: Video-ColBERT (VC in results; Reddy et al., 2025) and MMMORRF (MRF; Samuel et al., 2025), the state-of-the-art retrieval method on MultiVENT 2.0. We generate articles using the top 5 videos.

Table 5 reports article generation results using MMMORRF (nDCG@5: 0.66) and Video-ColBERT (nDCG@5: 0.22). We observe a significant decrease in CAG performance in moving from oracle retrieval to the RAG setting. This failure falls on the aggregation module of CAG: the text-only aggregator LLM struggles to include information from *each* video summary, even for irrelevant videos. In such cases, we find that the aggregator usually partitions the lead section into distinct topics instead of writing about the event covered by the (relevant) majority of retrieved videos.

6 Conclusion

In this paper we introduce the difficult task of automatically generating Wikipedia-style articles based on multiple videos about real-world events. We collect and release WIKIVIDEO, a benchmark of high-quality, expert-written articles grounded in diverse videos, ranging from amateur footage to professional news coverage, which are densely annotated for multimodal support of the articles' claims. Further, since existing systems for video-based summarization tasks are memory-intensive and overly focused on low-level video descriptions, we introduce **C**ollaborative **A**rticle **G**eneration (CAG)—a strong baseline for our task that leverages elements of relevance feedback and test-time scaling to iteratively construct *high-level* event-centric summaries. Our experiments demonstrate the effectiveness of CAG compared to alternative baselines—both in oracle retrieval and RAG settings. While CAG takes a significant step toward addressing the above limitations, future work remains in: (1) efficient multi-video inference (single inference step), (2) effectively integrating audio signal into the article generation process, (3) training VLMs to perform high-level inference, (4) connecting visual signals to parametric knowledge, and (5) improving video retrieval for RAG performance.

[7] Appendix E for qualitative comparison between CAG and LLM.

[8] A longer discussion of audio including is in Appendix F

Limitations

Parametric Knowledge Models trained on Wikipedia (Barham et al., 2023; Cheng et al., 2024) will be able to artificially inflate scores as quoting from a Wikipedia article leads to high scores across all metrics than human annotators (see Appendix H). We suggest reporting results on each subset (WIKIVIDEO-24 and WIKIVIDEO-25) to help with this limitation. However, when new models come out trained on Wikipedia articles from 2025, the data will again need to be updated.

Computational Costs Multi-video inference is an important problem to solve and a limiting factor of our method. To run inference on the 72B version of CAG, it requires 8 80GB A100s and this is only for *single video* inference. For multi-video this will continue to compound the computational costs and future work should be dedicated to decreasing these costs.

References

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

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. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Samuel Barham, Chandler May, and Benjamin Van Durme. 2025. Megawika 2: A more comprehensive multilingual collection of articles and their sources. *Preprint*, arXiv:2508.03828.

Samuel Barham, Orion Weller, Michelle Yuan, Kenton Murray, Mahsa Yarmohammadi, Zhengping Jiang, Siddharth Vashishtha, Alexander Martin, Anqi Liu, Aaron Steven White, Jordan Boyd-Graber, and Benjamin Van Durme. 2023. Megawika: Millions of reports and their sources across 50 diverse languages. *Preprint*, arXiv:2307.07049.

David Chen and William Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA. Association for Computational Linguistics.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2024. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36.

Yunmo Chen, William Gantt, Tongfei Chen, Aaron White, and Benjamin Van Durme. 2023a. A unified view of evaluation metrics for structured prediction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12868–12882, Singapore. Association for Computational Linguistics.

Yunmo Chen, William Gantt, Weiwei Gu, Tongfei Chen, Aaron White, and Benjamin Van Durme. 2023b. Iterative document-level information extraction via imitation learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1858–1874, Dubrovnik, Croatia. Association for Computational Linguistics.

Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2024. Dated data: Tracing knowledge cutoffs in large language models. In *First Conference on Language Modeling*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Xinya Du, Alexander Rush, and Claire Cardie. 2021. GRIT: Generative role-filler transformers for document-level event entity extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 634–644, Online. Association for Computational Linguistics.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

William Gantt, Alexander Martin, Pavlo Kuchmiichuk, and Aaron Steven White. 2024. Event-keyed summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7333–7345, Miami, Florida, USA. Association for Computational Linguistics.

Anisha Gunjal and Greg Durrett. 2024. Molecular facts: Desiderata for decontextualization in LLM fact verification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3751–3768, Miami, Florida, USA. Association for Computational Linguistics.

Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating summaries from user videos. In *Computer Vision – ECCV 2014*, pages 505–520, Cham. Springer International Publishing.

Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. 2023. Align and Attend: Multimodal Summarization with Dual Contrastive Losses . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14867–14878, Los Alamitos, CA, USA. IEEE Computer Society.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Qisheng Hu, Quanyu Long, and Wenya Wang. 2025. Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance? *Preprint*, arXiv:2411.02400.

Hang Hua, Yunlong Tang, Chenliang Xu, and Jiebo Luo. 2024. V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning. *Preprint*, arXiv:2404.12353.

Kung-Hsiang Huang, Philippe Laban, Alexander Fabbri, Prafulla Kumar Choubey, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2024. Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 570–593, Mexico City, Mexico. Association for Computational Linguistics.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *Preprint*, arXiv:2503.06749.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. *Preprint*, arXiv:1704.04497.

William Jurayj, Jeffrey Cheng, and Benjamin Van Durme. 2025. Is that your final answer? test-time scaling improves selective question answering. *Preprint*, arXiv:2502.13962.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

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. 2024. Multivent 2.0: A massive multilingual benchmark for event-centric video retrieval. *Preprint*, arXiv:2410.11619.

Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, and Luca Soldaini åand Eugene Yang. 2024. Overview of the trec 2023 neuclir track. *Preprint*, arXiv:2404.08071.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium. Association for Computational Linguistics.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19108–19118.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. 2025. Videochat-flash: Hierarchical compression for long-context video modeling. *Preprint*, arXiv:2501.00574.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jingyang Lin, Hang Hua, Ming Chen, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Jiebo Luo. 2024. Videoxum: Cross-modal visual and textural summarization of videos. *IEEE Transactions on Multimedia*, 26:5548–5560.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018a. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam M. Shazeer. 2018b. Generating wikipedia by summarizing long sequences. *ArXiv*, abs/1801.10198.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence

RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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*.

Arun Reddy, Alexander Martin, Eugene Yang, Andrew Yates, Kate Sanders, Kenton Murray, Reno Kriz, Celso M. de Melo, Benjamin Van Durme, and Rama Chellappa. 2025. Video-colbert: Contextualized late interaction for text-to-video retrieval. *Preprint*, arXiv:2503.19009.

Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. 2025. Videorag: Retrieval-augmented generation with extreme long-context videos. *Preprint*, arXiv:2502.01549.

J. J. Rocchio. 1971. Relevance feedback in information retrieval.

Saron Samuel, Dan DeGenaro, Jimena Guallar-Blasco, Kate Sanders, Oluwaseun Eisape, Arun Reddy, Alexander Martin, Andrew Yates, Eugene Yang, Cameron Carpenter, David Etter, Efsun Kayi, Matthew Wiesner, Kenton Murray, and Reno Kriz. 2025. Mmmorrf: Multimodal multilingual modularized reciprocal rank fusion. *Preprint*, arXiv:2503.20698.

Kate Sanders, David Etter, Reno Kriz, and Benjamin Van Durme. 2023. MultiVENT: Multilingual videos of events and aligned natural text. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Christina Sauper and Regina Barzilay. 2009. Automatically generating Wikipedia articles: A structure-aware approach. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 208–216, Suntec, Singapore. Association for Computational Linguistics.

Yijia Shao, Yucheng Jiang, Theodore A. Kanell, Peter Xu, Omar Khattab, and Monica S. Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. *Preprint*, arXiv:2402.14207.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. Tvsum: Summarizing web videos using titles. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5179–5187.

Neha Srikanth and Rachel Rudinger. 2025. Nli under the microscope: What atomic hypothesis decomposition reveals. *Preprint*, arXiv:2502.08080.

Sai Vallurupalli, Sayontan Ghosh, Katrin Erk, Niranjan Balasubramanian, and Francis Ferraro. 2022. POQue: Asking participant-specific outcome questions for a deeper understanding of complex events. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8674–8697, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Siddharth Vashishtha, Alexander Martin, William Gantt, Benjamin Van Durme, and Aaron White. 2024. FAMuS: Frames across multiple sources. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8250–8273, Mexico City, Mexico. Association for Computational Linguistics.

William Walden, Pavlo Kuchmiichuk, Alexander Martin, Chihsheng Jin, Angela Cao, Claire Sun, Curisia Allen, and Aaron Steven White. 2024. Cross-document event-keyed summarization. *Preprint*, arXiv:2410.14795.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2020. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. *Preprint*, arXiv:1904.03493.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. 2025. Internvideo2.5: Empowering video mllms with long and rich context modeling. *Preprint*, arXiv:2501.12386.

Miriam Wanner, Benjamin Van Durme, and Mark Dredze. 2024a. Dndscore: Decontextualization and decomposition for factuality verification in long-form text generation. *Preprint*, arXiv:2412.13175.

Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. 2024b. A closer look at claim decomposition. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 153–175, Mexico City, Mexico. Association for Computational Linguistics.

Orion Weller, Kathryn Ricci, Eugene Yang, Andrew Yates, Dawn Lawrie, and Benjamin Van Durme. 2025. Rank1: Test-time compute for reranking in information retrieval. *Preprint*, arXiv:2502.18418.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296.

Zhongyu Yang, Jun Chen, Dannong Xu, Junjie Fei, Xiaoqian Shen, Liangbing Zhao, Chun-Mei Feng, and Mohamed Elhoseiny. 2025. Wikiautogen: Towards multi-modal wikipedia-style article generation. *Preprint*, arXiv:2503.19065.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: a dataset for understanding complex web videos via question answering. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024a. Video instruction tuning with synthetic data. *Preprint*, arXiv:2410.02713.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun MA, Ziwei Liu, and Chunyuan Li. 2024b. Video instruction tuning with synthetic data.

Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J. Corso, and Marcus Rohrbach. 2019. Grounded video description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fangwei Zhu, Shangqing Tu, Jiaxin Shi, Juanzi Li, Lei Hou, and Tong Cui. 2021. TWAG: A topic-guided Wikipedia abstract generator. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4623–4635, Online. Association for Computational Linguistics.

A Dataset Statistics

Table 6 contains additional statistics about the claims in WIKIVIDEO and Table 7 has additional statistics about the videos. Further examples of WIKIVIDEO articles can be found in Table 21 and Table 22. We also report the IAA metrics in Table 8.

Video-Grounded Claims	954
Audio-Grounded Claims	1299
A+V-Grounded Claims	674
Avg. Claims / Event	51.10
Avg. Audio Claims / Event	24.98
Avg. Video Claims / Event	18.35
Avg. OCR Claims / Event	28.53
Avg. Claims in All / Event	12.96

Table 6: Claim Statistics

Max Videos for a Topic	12
Avg. Videos / Event	7.49
Total Relevant Videos	398
RAG Video Data Lake	109K
Avg. Video Length (Relevant)	79.57s
Avg. Video Length (RAG)	145s
Max Relevant Video Length	586.26s
Min Relevant Video Length	4.55s

Table 7: Video Statistics

B Data Collection

This appendix discusses the annotation process for WIKIVIDEO in greater detail. Beyond topic selection, recall that the annotation process includes claim decomposition, subclaim rewriting, subclaim grounding, and article rewriting.

B.1 Claim Decomposition

Much recent work has studied the appropriate granularity of subclaims in a claim decomposition and the applications of such decompositions to natural language inference (NLI) and claim verification

Modality	α
Video	0.446
Audio	0.780
OCR	0.722
None	0.682
Overall	0.767

Table 8: Krippendorff's α for the claim grounding agreement of each modality. Each judgment reflects a binary decision about whether a given claim is supported by a given modality (or **None** of the modalities).

(Min et al., 2023; Gunjal and Durrett, 2024; Wanner et al., 2024b,a; Hu et al., 2025; Srikanth and Rudinger, 2025). Our own claim decomposition method is most similar to those of (Gunjal and Durrett, 2024) and (Wanner et al., 2024a) in that we *decontextualize* subclaims—insert elided or abstracted context (e.g. by substituting pronouns with named entities)—up to the point that no further verification of extracted facts is required. (An example of one of our claim decomposition prompts can be found in Figure 4). Although not a rigourous form of decomposition, this notion is straightforward to apply during annotation. Additionally for simplicity, we treat dates as named entities and do not decompose them beyond their mention as given in the text. For example, some methods, like (Wanner et al., 2024b), may decompose the claim "The event occurred on 15 April 2019" as:

- The event occurred on the 15th
- The event occurred in April
- The event occurred in 2019

However, this produces an additional burden on our downstream annotations requiring the annotator to verify all 3 concepts. While it is possible to fail to recover finer-grained information that may be attested in the video—e.g. *the event occurred in 2019*—such cases are rare and this decision substantially reduces the amount of labor required for subclaim rewriting and grounding.

B.2 Subclaim Rewriting

We take the subclaims decomposed by Qwen2.5 32B and correct them manually, with three of the authors serving as annotators. Figure 5 shows the interface that the annotators used to rewrite, add, and remove claims, and Figure 6 shows the instructions provided to annotators.

B.3 Subclaim Grounding

Once the final set of claims is completed by the human annotator, the same annotators ground the claims in the video content. We provide the instructions for this process in Figure 8 and the protocol for this populated with an example sentence worth of claims and video from the Notre-Dame fire in Figure 7.

Note that this grounding annotation was initially attempted via Amazon Mechanical Turk. However, despite several iterations of annotation instructions, even Turkers with additional Masters qualifications struggled to consistently ground claims in videos. We suspect this is in part due to the amount of domain expertise and world knowledge required to properly ground claims, and partially because raw real-time videos of events are inherently ambiguous. Thus, the decision was made to leverage domain experts in order to ensure consistent and high-quality annotations.

B.4 Article Rewriting

The last stage in the annotation process involves rewriting the original Wikipedia articles based on the grounded claims. The instructions for this are provided in Figure 9.

C Relevance and RePrompt

In this section, we include the prompt used with the reasoning model (Figure 10) and examples of real queries provided to the VideoLLM (Figure 11). We note that we had to add force the model to produce the prefix "Describe the video in detail and focus on ..." or else sometimes the reasoner would not follow our instructions and produce a new query.

D Article Generation

In this section we present the prompt for article synthesis (Figure 12)

E Qualitative Differences Between Outputs

Qualitative Results for Model Variations In this section we present additional qualitative examples from the different generation methods. We present these for the Notre-Dame fire query against the reference article (Table 9, Table 10, Table 11).

Qualitative Results for Audio vs. Video We also present the qualitative differences between

Instructions: - You are given a paragraph, and one sentence from the paragraph to decompose - You must decompose this into a set of claims - You must decompose this into a JSON format ["claim": "...", "claim": "...", ...]
PARAGRAPH: On 15 April 2019, just before 18:20 CEST, a structural fire broke out in the roof space of Notre-Dame de Paris, a medieval Catholic cathedral in Paris, France. By the time the fire was extinguished, the cathedralš wooden spire (flèche) had collapsed, most of the wooden roof had been destroyed, and the cathedralš upper walls were severely damaged. Extensive damage to the interior was prevented by the vaulted stone ceiling, which largely contained the burning roof as it collapsed. Many works of art and religious relics were moved to safety, but others suffered smoke damage, and some of the exterior art was damaged or destroyed. The cathedralš altar, two pipe organs, and three 13th-century rose windows suffered little or no damage. Three emergency workers were injured. The fire contaminated the site and nearby areas of Paris with toxic dust and lead. Notre-Dame did not hold a Christmas Mass in 2019, for the first time since 1803. Investigators in 2020 believed the fire to have been "started by either a cigarette or a short circuit in the electrical system".
SENTENCE: On 15 April 2019, just before 18:20 CEST, a structural fire broke out in the roof space of Notre-Dame de Paris, a medieval Catholic cathedral in Paris, France.
DECOMPOSITION: ["claim": "A structural fire broke out", "claim": "The fire broke out on 15 April 2019", "claim": "The fire broke out just before 18:20 CEST", "claim": "The fire broke out in the roof space", "claim": "Notre-Dame de Paris is a medieval Catholic cathedral", "claim": "Notre-Dame de Paris is located in Paris, France"]
PARAGRAPH: The 2022 United States Senate election in Georgia was held on November 8, 2022, to elect a member of the U.S. Senate to represent the state of Georgia. Incumbent Democratic senator Raphael Warnock won his first full term in office, defeating Republican former football player Herschel Walker. Under Georgia"s two-round system, Warnock was re-elected in a runoff election on December 6 after neither candidate received over 50% of the vote on November 8. Warnock"s win was the only statewide victory for Democrats in Georgia in 2022, as Republicans swept all other races.
SENTENCE: Under Georgia"s two-round system, Warnock was re-elected in a runoff election on December 6 after neither candidate received over 50
DECOMPOSITION: ["claim": "Georgia has a two-round election system", "claim": "The runoff election is part of the two-round election system in Georgia", "claim": "A run off election occurs when no candidate receives 50% of the vote", "claim": "Neither candidate received more than 50% of the vote in the first election", "claim": "The runoff election took place on December 6", "claim": "Warnock won the runoff election", "claim": "Warnock was the incumbent candidate", "claim": "The first election took place on November 8"]
PARAGRAPH: Hurricane Irma was an extremely powerful Cape Verde hurricane that caused widespread destruction across its path in early September 2017. Irma was the first Category 5 hurricane to strike the Leeward Islands on record, followed by Maria two weeks later. At the time, it was considered the most powerful hurricane on record in the open Atlantic region, outside of the Caribbean Sea and Gulf of Mexico, until it was surpassed by Hurricane Dorian two years later. It was also the third-strongest Atlantic hurricane at landfall ever recorded, just behind the 1935 Labor Day Hurricane and Dorian. The ninth named storm, fourth hurricane, second major hurricane, and first Category 5 hurricane of the extremely active 2017 Atlantic hurricane season, Irma caused widespread and catastrophic damage throughout its long lifetime, particularly in the northeastern Caribbean and the Florida Keys. It was also the most intense hurricane to strike the continental United States since Katrina in 2005, the first major hurricane to make landfall in Florida since Wilma in the same year, and the first Category 4 hurricane to strike the state since Charley in 2004. The word Irmageddon was coined soon after the hurricane to describe the damage caused by the hurricane.
SENTENCE: The ninth named storm, fourth hurricane, second major hurricane, and first Category 5 hurricane of the extremely active 2017 Atlantic hurricane season, Irma caused widespread and catastrophic damage throughout its long lifetime, particularly in the northeastern Caribbean and the Florida Keys.
DECOMPOSITION: ["claim": "Irma was the ninth named storm of the 2017 Atlantic hurricane season", "claim": "Irma was the fourth hurricane of the 2017 Atlantic hurricane season", "claim": "Irma was the second major hurricane of the 2017 Atlantic hurricane season", "claim": "Irma was the first Category 5 hurricane of the 2017 Atlantic hurricane season", "claim": "The 2017 Atlantic hurricane season was extremely active", "claim": "Irma caused widespread damage", "claim": "Irma caused catastrophic damage", "claim": "Irma's damage was particularly severe in the northeastern Caribbean", "claim": "Irma's damage was particularly severe in the Florida Keys", "claim": "Irma had a long lifetime", "claim": "Irma occurred during the 2017 Atlantic hurricane season"]
PARAGRAPH: On November 30, 2018, at 8:29 a.m. AKST (17:29 UTC), a magnitude 7.1 earthquake hit Anchorage in South Central Alaska. The earthquake's epicenter was near Point Mackenzie, about north of Anchorage, and occurred at a depth of . It was followed six minutes later by a magnitude 5.7 aftershock centered north-northwest of the municipality. The earthquake could be felt as far away as Fairbanks.
SENTENCE: On November 30, 2018, at 8:29 a.m. AKST (17:29 UTC), a magnitude 7.1 earthquake hit Anchorage in South Central Alaska.
DECOMPOSITION: ["claim": "The earthquake occurred", "claim": "The earthquake occurred on November 30, 2018", "claim": "The earthquake occurred at 8:29 a.m. AKST", "claim": "The earthquake hit Anchorage", "claim": "The earthquake hit South Central Alaska", "claim": "The earthquake had a magnitude of 7.1"]
PARAGRAPH: Pokémon Go (stylized as Pokémon GO) is a 2016 augmented reality (AR) mobile game, part of the Pokémon franchise, developed and published by Niantic in collaboration with Nintendo and The Pokémon Company for iOS and Android devices. It uses mobile devices with GPS to locate, capture, train, and battle virtual Pokémon, which appear as if they are in the player's real-world location. The game is free-to-play; it uses a freemium business model combined with local advertising and supports in-app purchases for additional in-game items. The game launched with around 150 species of Pokémon, which had increased to around 700 by 2021.
SENTENCE: Pokémon Go (stylized as Pokémon GO) is a 2016 augmented reality (AR) mobile game, part of the Pokémon franchise, developed and published by Niantic in collaboration with Nintendo and The Pokémon Company for iOS and Android devices.
DECOMPOSITION: ["claim": "Pokémon Go is a mobile game", "claim": "Pokémon Go is an augmented reality (AR) game", "claim": "Pokémon Go was released in 2016", "claim": "Pokémon Go is part of the Pokémon franchise", "claim": "Pokémon Go was developed by Niantic", "claim": "Pokémon Go was published by Niantic", "claim": "Niantic collaborated with Nintendo to develop Pokémon Go", "claim": "Niantic collaborated with Nintendo to publish Pokémon Go", "claim": "Niantic collaborated with The Pokémon Company to develop Pokémon Go", "claim": "Niantic collaborated with The Pokémon Company to publish Pokémon Go", "claim": "Pokémon Go was developed for iOS devices", "claim": "Pokémon Go was published for iOS devices", "claim": "Pokémon Go was developed for Android devices", "claim": "Pokémon Go was published for Android devices"]

PARAGRAPH [paragraph] SENTENCE [sentence] DECOMPOSITION:

Figure 4: Prompt For Qwen 2.5 32B Claim Decomposition

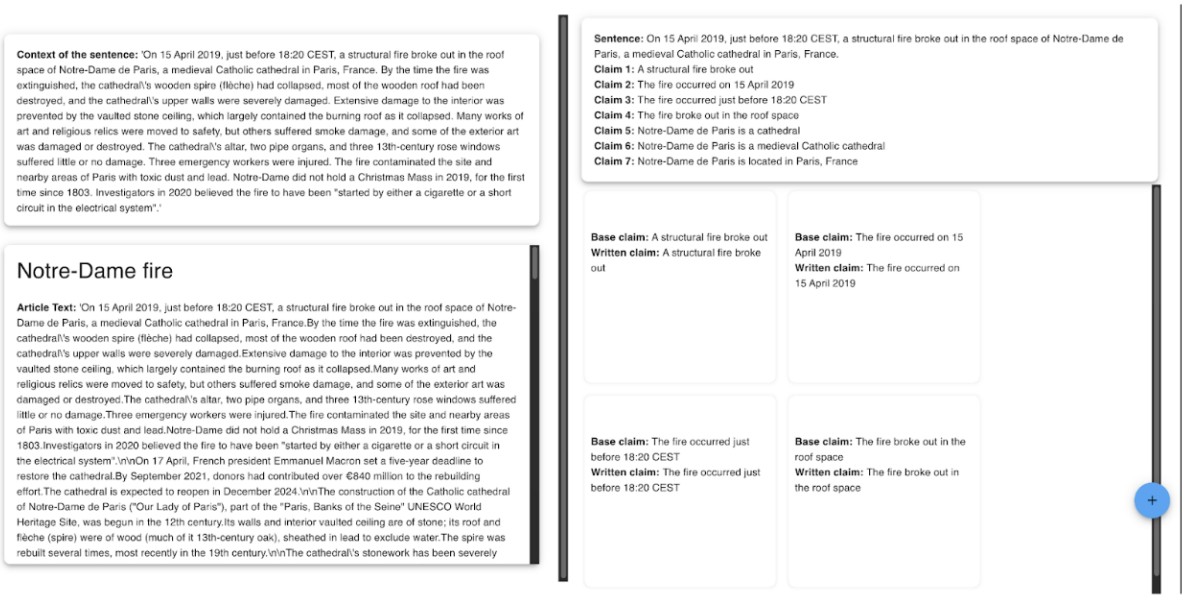

Figure 5: The annotation interface for our subclaim grounding task. In this protocol, the left hand side is both versions of the Wikipedia context. The top context is the paragraph a sentence comes from and the bottom context is the lead section of the Wikipedia article. On the right hand side is the sentence to be decomposed and its claims. The claims from Qwen32B are prepopulated in the protocol and the rewriters edit them.

LLaVA-Video	VAST	InternVideo2.5	QwenVL2.5	Reference
The video begins with a view of a historic cathedral engulfed in flames, with thick smoke billowing into the sky. The cathedral's spire is prominently visible, and the surrounding buildings are also affected by the fire. The scene is dramatic, with the fire raging intensely and the smoke creating a dark cloud over the area. The video then transitions to a closer view of the cathedral's spire, which is now almost entirely consumed by the fire. The flames are bright and fierce, illuminating the intricate details of the spire. The smoke continues to rise, obscuring parts of the structure. The video progresses to show the spire collapsing under the heat, with the flames still burning fiercely. The final frames focus on the aftermath of the collapse, with the remains of the spire and the surrounding area still ablaze. The video concludes with a wider shot of the cathedral, showing the extent of the damage and the ongoing fire.	fire is burning on a tall tower that has been burning	The video captures the catastrophic fire engulfing the Notre Dame Cathedral in Paris. The iconic structure is ablaze, with flames and smoke billowing from its spire and roof. The fire's orange glow contrasts with the grey sky, while thick smoke rises into the air. The surrounding environment appears calm, with no visible people or firefighters in the immediate vicinity of the blaze. The cathedral's gothic architecture is partially visible, with its skeletal framework exposed by the fire. The scene is somber, highlighting the loss of a historical landmark.	The video depicts a catastrophic fire engulfing a historic cathedral, likely Notre-Dame in Paris, France. The flames are intense and widespread, with thick smoke billowing into the sky. The iconic spire of the cathedral is visibly on fire, with the structure glowing red from the heat. As the fire rages, the spire begins to weaken and eventually collapses, sending a cascade of debris downward. The surrounding buildings and trees are partially obscured by the smoke, emphasizing the scale and severity of the blaze. The scene is both dramatic and tragic, capturing a moment of significant destruction to a beloved landmark.	A structural fire broke out in the roof space of the Notre-Dame de Paris, a medieval Catholic cathedral in Paris, France. By the time the fire was extinguished, the cathedral's wooden spire, called a fléche, collapsed, most of the wooden roof had been destroyed, and the cathedral's upper walls were severely damaged. Extensive damage to the cathedral's interior was prevented by the vaulted stone ceiling, which largely contained the burning roof as it collapsed. However, some works of art and religious relics suffered smoke damage and some exterior art was damaged or destroyed. The cathedral's altar and three 13th-century rose windows suffered little or no damage. There were emergency workers present during the fire. The fire contaminated the site and nearby areas of Paris with toxic dust. French president Emmanuel Macron set a deadline for the restoration of the cathedral.

Table 9: Qualitative Results for VideoLLMs on the Notre-Dame Fire query. The results of the table are from video ID: 1117864206475218944. **Bold text represents model hallucinations.** This is any information included in the article that is not in any of the videos.

Claim Rewriting Annotation
In this task, you will be shown an excerpt from Wikipedia, a sentence, and a set of claims associated with that sentence. Given this information, you will rewrite the set of claims into a new set of “cleaner” claims. To create the final set of claims, you will deal with common issues like splitting a claim into two or more claims, adding a missing claim, or removing a duplicated claim.
Do not remove duplicates you remember from a previous sentence of the document. Treat each sentence as a unique instance without recalling what you had annotated previously.
Annotation Protocol
The annotation protocol has 4 main sections. The R.H.S. is what is most relevant to annotators. It contains the current sentence and the claims for the sentences as well as editable cards with each claim from the sentence in them. (Sometimes not all of them are prepopulated, so if any are missing just hit the + button). The L.H.S. has the context that the sentences are taken from. The top context is the paragraph where the current sentence comes from and below is the lead section of the wikipedia article that the paragraph comes from.
Editing a claim
To edit a claim, you will click on the card that the claim is in. The claim will include above it the original claim and the textbox will allow you to delete, edit, and rewrite claims.
Adding extra claims Sometimes claims are missed by the decomposer. To fix these press the + button( ) in the right side of the interface. This will put a new claim into the interface. Note that these claims won’t include base claims in the interface because the system did not predict the claim. This will NOT impact the annotation or future use of the claim.
Error Types Here are some examples of errors that you might encounter when doing the annotations.
Under Decomposition Under decomposition is when a claim includes multiple pieces of information that could be split into two or more atomic facts.
Under Decomposition Example (loaded claim): Input: On 15 April 2019, just before 18:20 CEST, a structural fire broke out in the roof space of Notre-Dame de Paris, a medieval Catholic cathedral in Paris, France. Incorrect Decomposition: A structural fire broke out The fire occurred on 15 April 2019 The fire occurred just before 18:20 CEST The fire broke out in the roof space Notre-Dame de Paris is a cathedral Notre-Dame de Paris is a medieval Catholic cathedral Notre-Dame de Paris is located in Paris, France Correct Decomposition: A structural fire broke out The fire occurred on 15 April 2019 The fire occurred just before 18:20 CEST The fire broke out in the roof space Notre-Dame de Paris is a cathedral Notre-Dame de Paris is a medieval Catholic cathedral Notre-Dame de Paris is located in Paris Notre-Dame de Paris is located in France Reasoning: Original claim 7 had 2 pieces of information in it: The location of Paris and The location of France. While this may seem intuitive that Paris is in France, from an evaluation perspective, it is better to have two distinct claims to verify. See another brief example below:
Incorrect Decomposition: The event happened in Seoul, South Korea Correct Decomposition: The event happened in Seoul The event happened in South Korea Reasoning: When considering the evaluation of where the event happened, it’s better to split the claim into the two locations: Seoul and South Korea, so that you can evaluate against systems that say only South Korea (if it happened in other locations in SK) or against systems that only state the city.
Hallucinated Decomposition This is the addition of a claim that isn’t supported by the sentence.
Input: The 2016 World Short Track Speed Skating Championships took place from 11 to 13 March 2016 in Seoul, South Korea. Incorrect Decomposition: The event took place The event was from 11 to 13 March 2016 The event happened in Seoul The event happened in South Korea The event was the 41st speed skating championship Correct Decomposition: The event took place The event was from 11 to 13 March 2016 The event happened in Seoul The event happened in South Korea Reasoning: 5 is factual, but it is not supported by the sentence.
Ambiguity Only resolve ambiguity if the claims make it difficult to disambiguate between entities.
Input: The 2016 World Short Track Speed Skating Championships took place from 11 to 13 March 2016 in Seoul, South Korea. Correct Decomposition: The event took place The event was from 11 to 13 March 2016 The event happened in Seoul The event happened in South Korea Reasoning: Only one event in the sentence. No need to disambiguate.
Input: Due to Imran Khan’s criticism of Macron’s comments on Islam, French authorities cancelled the visas of 183 Pakistani citizens and deported 118 from the country. Incorrect Decomposition: They cancelled the visas of 183 Pakistani citizens. They deported 118 Pakistani citizens from the country. He criticized Macron’s comments on Islam Correct Decomposition: French authorities cancelled the visas of 183 Pakistani citizens. French authorities deported 118 Pakistani citizens from the country. Imran Khan criticized Macron’s comments on Islam Reasoning: In this scenario, it’s better to disambiguate the references to the named entities because they could be Imran, Macron, or the French authorities.
Current notes / edge cases during test evaluation Go over 2022 Senate Election annotations.
Sentence: The National Tsunami Warning Center—itself located inside the quake zone, in Palmer, Alaska, northeast of Anchorage—issued tsunami warnings for nearby coastal areas, including Cook Inlet and the Kenai Peninsula, but they were lifted shortly after. Claim 1: The National Tsunami Warning Center issued warnings Claim 2: The warnings were for nearby coastal areas Claim 3: The warnings included Cook Inlet Claim 4: The warnings included the Kenai Peninsula Claim 5: The warnings were lifted shortly after issuance Claim 6: Palmer is located in Alaska Claim 7: Palmer is northeast of Anchorage Claim 8: The National Tsunami Warning Center is located in Palmer Claim 9: The National Tsunami Warning Center is inside the quake zone Claim 10: Cook Inlet is a coastal area Claim 11: The Kenai Peninsula is a coastal area
6,7,8 as claims related to the location of NTWC or Palmer. Perspective matters probably. I would may rewrite these to be all about the NTWC location.
Sentence: Notre-Dame did not hold a Christmas Mass in 2019, for the first time since 1803. Claim 1: Notre-Dame did not hold a Christmas Mass in 2019 Claim 2: Notre-Dame did not hold a Christmas Mass was in 1803 Claim 3: Notre-Dame held a Christmas Mass every year between 1803 and 2019
Sentence: Investigators in 2020 believed the fire to have been "started by either a cigarette or a short circuit in the electrical system". Claim 1: The investigators believed the fire was started Claim 2: The investigators identified two possible causes for the fire Claim 3: One possible cause was a cigarette Claim 4: Another possible cause was a short circuit in the electrical system Claim 5: The investigation took place in 2020
New: There was an investigation into the cause of the fire The investigators identified two possible causes for the fire The fire was possibly started by a cigarette The fire was possibly started by a short circuit in the electrical system
The investigation took place in 2020 (this might not be factual) How to incorporate the date? In 2020 investigators believed the fire was started.
original: 5 [’The investigators believed the fire was started’, ’The investigators identified two possible causes for the fire’, ’One possible cause was a cigarette’, ’Another possible cause was a short circuit in the electrical system’, ’The investigation took place in 2020’] 2958: 4 [’The investigators identified two possible causes for the fire’, ’A possible cause was a cigarette’, ’A possible cause was a short circuit in the electrical system’, ’The investigation took place in 2020’] 2959: 6 [’The investigators believed the fire was started’, ’The investigators identified two possible causes for the fire’, ’The fire was possibly started by a cigarette’, ’The fire was possibly started by a short circuit.’, ’The investigation took place in 2020’, ’The possible short circuit occurred in the electrical system’] 2960: 5 [’There was an investigation into the cause of the fire.’, ’Investigators identified two possible causes for the fire’, ’One possible cause was a cigarette’, ’One possible cause was a short circuit in the electrical system’, ’An investigation took place in 2020’]

Figure 6: Annotation Instructions for Claim Rewriting

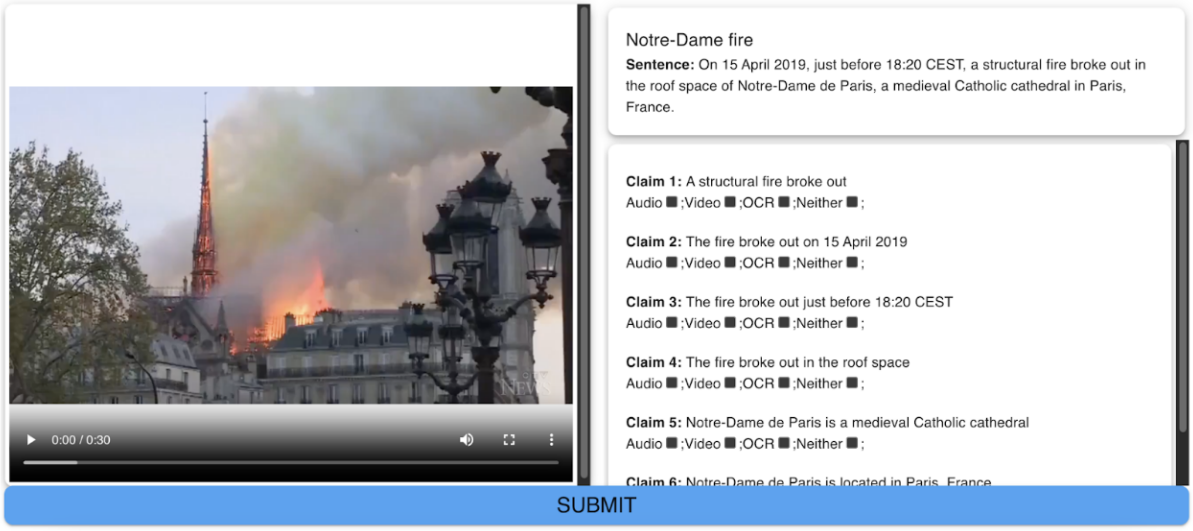

Figure 7: The claim grounding protocol. This protocol has the video on the left hand side and a sentence and its claims on the right. Each claim has 4 buttons which the annotator can select for the modality (or none) that support the claim.

LLaVA-Video	VAST	InternVideo2.5	QwenVL2.5	Reference
On April 15, 2019, a devastating fire broke out at Notre-Dame Cathedral in Paris, one of the world's most famous and historic Gothic landmarks. The blaze, which began in the attic, rapidly spread, causing significant damage to the cathedral's wooden frame and leading to the collapse of its spire. The fire raged for several hours, with thick plumes of smoke and flames visible from across the city. Despite the extensive damage, firefighters managed to save the cathedral's main structure, including the two bell towers. **The incident drew widespread international attention and prompted pledges of support for the reconstruction efforts from global leaders and private donors.** President Emmanuel Macron vowed that the cathedral would be restored, **and initial assessments indicated that the restoration process could take years to complete. The fire raised questions about the safety and maintenance of historical sites and the methods used to protect them from such disasters.**	A major fire engulfed a prominent building in a French city, drawing a crowd of onlookers who gathered to witness the event. Flames and thick smoke were visible from various parts of the structure, including its roof and upper levels. The fire, which broke out in the middle of the building, **also affected a nearby church, causing significant damage.** Emergency services responded to the scene, while local residents and visitors stopped to observe and document the incident, some taking photographs and discussing the unfolding situation. **The cause of the fire and the extent of the damage remained under investigation.**	A catastrophic fire engulfed the Notre-Dame Cathedral in Paris **on April 15, 2019**, causing extensive damage to the historic structure. Flames and thick smoke billowed from the roof and spire, which partially collapsed during the blaze. **The fire, which began in the attic, spread rapidly, threatening the integrity of the entire building.** Despite the intense heat and the structural damage, the main facade and two bell towers remained standing. Emergency responders, including hundreds of firefighters, worked tirelessly throughout the night to control the fire. The event drew large crowds of onlookers and prompted a widespread outpouring of grief and solidarity, both locally and internationally. **The cause of the fire was initially suspected to be linked to renovation work being carried out on the cathedral.** The aftermath revealed a charred interior and significant structural damage, though some elements, such as the stained glass windows, survived. Restoration efforts were immediately pledged by the French government **and private donors.**	**On April 15, 2019**, a catastrophic fire broke out at Notre-Dame Cathedral in Paris, France, causing extensive damage to the historic landmark. The blaze, which erupted in the afternoon, rapidly spread through the cathedral's roof and spire, leading to the collapse of the iconic spire and significant portions of the roof. Thick plumes of smoke billowed into the sky, visible from across the city, as firefighters worked tirelessly to contain the inferno. The fire drew a large crowd of concerned onlookers, including Parisians and tourists, who gathered to watch the unfolding disaster, many expressing their shock and sorrow. Despite the extensive damage, the structural integrity of the cathedral's main walls and towers was preserved, and key religious artifacts, including the cross and some statues, remained intact. President Emmanuel Macron pledged to rebuild the cathedral, **which is a UNESCO World Heritage site and a symbol of French heritage and culture. The cause of the fire was initially suspected to be linked to ongoing renovation work, and investigations were launched to determine the exact cause. The event highlighted the global significance of Notre-Dame and the profound impact of its partial destruction on both the local community and the international community.**	A structural fire broke out in the roof space of the Notre-Dame de Paris, a medieval Catholic cathedral in Paris, France. By the time the fire was extinguished, the cathedral's wooden spire, called a flèche, collapsed, most of the wooden roof had been destroyed, and the cathedral's upper walls were severely damaged. Extensive damage to the cathedral's interior was prevented by the vaulted stone ceiling, which largely contained the burning roof as it collapsed. However, some works of art and religious relics suffered smoke damage and some exterior art was damaged or destroyed. The cathedral's altar and three 13th-century rose windows suffered little or no damage. There were emergency workers present during the fire. The fire contaminated the site and nearby areas of Paris with toxic dust. French president Emmanuel Macron set a deadline for the restoration of the cathedral.

Table 10: Qualitative Results for Article Generation on the Notre-Dame Fire query. **Bold text represents model hallucinations.** This is any information included in the article that is not in any of the videos.

This task involves watching a video and deciding whether the video supports particular claims. A video supports a claim if it can be verified from the video's visual, audio, or text content.
Visual text (denoted "Text" in the interface and instructions) support refers to on-screen text that supports the claim. This may include (e.g.) text on a street sign, text scrolling across the screen in a news broadcast, text on someone's clothing, subtitles on the screen, etc.
Visual support refers to anything else on-screen besides text that supports the claim. This includes any action happening on-screen, still images, or other graphics (e.g. a map shown in a weather report) or animations.
Audio support refers to any sounds (e.g. sirens, gunshots) or speech (e.g. from a newscaster or from the person filming) that supports the claim.
Each claim may have only one of these types of support in a given clip, may have multiple types of support, or may not be supported at all. It is your job to determine which type(s) of support there are for each claim.
Note: Sometimes the clips may contain audio (speech) or visual text in a language that you do not speak. For these instances, do not try to annotate claims for the information in another language. Task Interface On the left side of the interface, you will see a video clip.
You can adjust the playback speed by clicking on the three vertical dots at the bottom right of the clip:
You can adjust the video's playback speed by clicking the three vertical dots on the bottom right:
Note: We do not suggest making the video full screen.
On the right side of the interface, you will see: The name of the event that the clip is about a sentence about that event a list of claims derived from that sentence
Each claim comes with three buttons: Audio (for audio support), Visual (for visual support), Text (for text support), and Neither (for none of the above):
For each claim, you should check all boxes corresponding to the types of support for the claim attested in the video:
If a claim has no support in the video, you should click only the Neither checkbox:
You are allowed to select any combination of the Audio, Visual, and Text checkboxes, but the interface will prevent you from selecting the Neither checkbox in combination with any of the other three.
When all claims have at least one button checked, the HIT will become ready for submission: the SUBMIT button will change from gray to blue and you then can submit the HIT:
Task Instructions Your job is to:
Watch the video clip For each claim, select all checkboxes corresponding to the type(s) of support it receives in the video (Audio, Text, or Visual) or Neither if it has no support
Note that you are allowed—and encouraged!—to rewatch the video if necessary in order to assess the types of support it provides for a particular claim.
At least one checkbox must be clicked for each claim before you can submit the HIT. What does it mean for a claim to be supported by visuals/audio/text?
You can conceive of each claim as both posing a question and providing an answer to that question.
For example, the claim Dominguez hit a home run can be understood to be: Implicitly asking the question: Did Dominguez hit a home run? Answering this question affirmatively: Yes, he did.
If the audio (sounds, speech), visuals (action, images, graphics), and on-screen text provides the same answer to the question as the claim—i.e. if the audio, visuals, or text provide evidence that the claim is true—then you should click the corresponding checkbox.
Alternatively, the clip may not provide an answer to the claim's question at all, or may even provide an answer that contradicts the answer implicit in the claim. In these cases, you should select only the Neither checkbox.
Note: in many cases, it will be difficult to say with absolute certainty whether a given claim is supported or not by the audio, visuals, or text. The standard for determining whether a claim has one of these types of support is not certainty, but rather high confidence, given the clip's contents. You will have to rely on your own judgment of what a normal person could confidently infer about the truth of the claim, having watched the video yourself. (See World Knowledge Support)
Audio Support Examples Audio support may come from speech or other sounds in the video clip. Below are two examples of cases in which a claim has audio support.
Example 1
Claim: Dominguez hit the first home run of his career. Video: https://www.youtube.com/watch?v=c6U4AnW4ohM Audio Evidence: The broadcaster in the video announces the homerun that Dominguez hit.
Example 2
Claim: Emergency responders went to the scene Video: https://www.youtube.com/watch?v=rVKwa4ZqQAA Audio Evidence: You can hear the sirens of the fire trucks in this video. Thus, audio support. Visual Support Examples Visual support can come from any type of non-text visual content, including any action happening in the video clip, or still images or graphics that are shown. We include two examples of visual support below.
Example 3
Claim: Dominguez hit a homerun Video: https://www.youtube.com/watch?v=c6U4AnW4ohM Explanation: In the video you can see Dominguez hit the homerun because he swings, hits the ball, it goes into the stands.
Example 4
Claim: The event took place in France Video: (shows the image of the Eiffel Tower below)
Explanation: The Eiffel Tower is an iconic landmark in Paris, France. Even if the video doesn't explicitly state that it is located in France, you can very confidently infer that the location is France based on this knowledge that the Eiffel Tower is in France.
Example 5
Claim: The hurricane hit florida Video: (shows the graphic of the Hurricane's trajectory below)
Explanation: Although the map doesn't explicitly say "Florida" you can see from the map that the Hurricane's path goes through Miami (and thus hits Flordia) Text Support Examples Text support can come from any type of text visible on screen. This can be text deployed on the screen in a news broadcast, street signs, posters in a protest, etc. Some examples of text support are shown below.
Example 6
Claim: The event took place on August 30th. Video: (includes the frame below) Explanation: The date is clearly displayed on the bottom left of the screen (30 Aug)
Example 7
Claim: City of London police were deployed. Video: (includes the frame below) Explanation: It's clear from the text on the officer's back in the frame below that the people shown are City of London police, who appear to be deployed.
Example 8
Claim: Mueller subpoenaed a former aide of Trump. Video: (includes the frame below) Explanation: The text shown in the "Breaking News" banner at the bottom of the screen clearly supports the claim.
World Knowledge Support
Claim: Texas borders Mexico
Explanation: Here in this example, nothing on the map shows the names of Mexico or Texas, but using common knowledge, you can see both the outline of Texas and Mexico. Thus,

Neither Support Examples Claim: Dominguez hit the last home run of his career Video: https://www.youtube.com/watch?v=c6U4AnW4ohM Explanation: This is about his first ever home run. You cannot tell that his career has ended or that he hasn't hit a homerun since this game.

Figure 8: Instructions for the Claim Grounding Protocol.

LLaVA-Video	VAST	InternVideo2.5	QwenVL2.5	Reference
On April 15, 2019, a massive fire broke out at Notre-Dame Cathedral in Paris, one of the most iconic and historically significant Gothic cathedrals **in the world**. The blaze, which **began in the attic** of the cathedral, rapidly spread, causing the collapse of the spire and significant damage to the roof and interior. Thick black smoke billowed into the sky, and the flames were visible from across the city. Despite the intensity of the fire, firefighters managed to control and eventually extinguish the blaze by early morning. The incident drew a large crowd of onlookers, many of whom gathered along the Seine River and near the cathedral, some singing hymns in response to the tragedy. President Emmanuel Macron pledged to rebuild the cathedral, **and restoration efforts were soon underway, with the goal of preserving and restoring this cherished landmark.** The fire caused extensive damage, including the loss of the spire and a significant portion of the roof, but the main structure and many valuable artifacts were saved. **The event highlighted the cathedral's enduring cultural and spiritual significance to both the French people and the global community.**	CAG was not run on VAST due to poor extraction capabilities.	A catastrophic fire engulfed the Notre Dame Cathedral in Paris **on April 15, 2019**, causing significant damage to the historic structure. Flames and thick smoke billowed from the roof and spire, which partially collapsed during the blaze. The fire, which spread rapidly, was visible from afar, drawing large crowds of onlookers who watched in shock and dismay. Emergency responders, including firefighters, were quickly deployed to the scene, working tirelessly to control the inferno and prevent further damage. Despite their efforts, the cathedral suffered extensive damage, with the roof and spire being among the most affected areas. The incident occurred during the evening, and the surrounding area was evacuated to ensure public safety. The aftermath revealed a charred and debris-filled interior, with the iconic spire and much of the roof destroyed. **The fire's cause was initially unknown, but investigations were launched to determine the origin and circumstances of the disaster.** The event sparked a global outpouring of support and grief, with many gathering to mourn the loss of this cultural and religious landmark.	**On April 15, 2019,** a catastrophic fire broke out at Notre-Dame Cathedral in Paris, France, causing extensive damage to the historic landmark. The fire, **which began in the attic** of the cathedral, rapidly spread, engulfing the roof and the iconic spire, which collapsed in a dramatic fall witnessed by a shocked crowd gathered outside. The cathedral, a masterpiece of Gothic architecture **dating back to the 12th century**, suffered significant structural damage, with the roof and spire completely destroyed. Despite the intense blaze, the main stone structure, including the two bell towers, remained standing. Firefighters worked tirelessly throughout the night to contain the fire, **using advanced equipment and techniques to prevent further damage**. **The event drew global attention, with people around the world expressing sorrow and solidarity**. In the aftermath, French President Emmanuel Macron pledged to rebuild the cathedral, **and a fundraising campaign was launched to support the restoration efforts. The fire raised questions about the safety and preservation of historic landmarks and highlighted the cultural and spiritual significance of Notre-Dame to both France and the international community.**	A structural fire broke out in the roof space of the Notre-Dame de Paris, a medieval Catholic cathedral in Paris, France. By the time the fire was extinguished, the cathedral's wooden spire, called a fléche, collapsed, most of the wooden roof had been destroyed, and the cathedral's upper walls were severely damaged. Extensive damage to the cathedral's interior was prevented by the vaulted stone ceiling, which largely contained the burning roof as it collapsed. However, some works of art and religious relics suffered smoke damage and some exterior art was damaged or destroyed. The cathedral's altar and three 13th-century rose windows suffered little or no damage. There were emergency workers present during the fire. The fire contaminated the site and nearby areas of Paris with toxic dust. French president Emmanuel Macron set a deadline for the restoration of the cathedral.

Table 11: Qualitative Results for CAG on the Notre-Dame Fire query. **Bold text represents model hallucinations.** This is any information included in the article that is not in any of the videos.

```
Your task is to write a new Wikipedia article to exclude the claims not found in the video content. You will be given a set of claims and the sentences
they come from on the L.H.S. of the protocol and your job will be to rewrite the article / sentences such that only the supported claims are presented
in the article. You should try to diversify your writing from the Wikipedia if possible without stepping too far away from the general “Wikipedia”
style.
```

Figure 9: Instructions for Article Rewriting

```
I am trying to find information about EVENT_QUERY. I will show you a video summary that might be related to the event. Based on the current summary,
can you think of a new query that might help me find more information about the event? Please write a new query that you think will help me find more
information about the event. DO NOT write anything except for the new query. If you think the current summary is sufficient, you can say ’no new
query.’ Otherwise, start your new query with ’Describe the video in detail and focus on’ Here is the video summary:
```

Figure 10: Prompt For Qwen32B Distilled R1.

video only and video+audio CAG results in Table 12, Table 13, and Table 14. In these tables you can see that the articles produced using transcripts are shorter than the articles based only on video content. Numerically, the average length of a video-only article is 206.36 tokens and the average length of a article with transcript provided is 163.90 tokens.

Qualitative Results for 2025 Subset We also present the qualitative differences between CAG and LLM responses for the 2025 subset of WIKIVIDEO. The LLM is prompted with "Generate a Wikipedia article about [event]." Table 15 shows the comparison in these results. However, the LLM results are poor, filled with ungrounded and not factual information. In the main text, it's noted that the Rouge and BertScores show similar performance of the LLM to CAG-R. However, these qualitative examples show that the LLM is only matching the style of Wikipedia articles and not generating meaningful and factual text.

F Incorporating Audio

Articles in WIKIVIDEO have many claims grounded partly or only in videos' audio signal (Table 1). Here, we consider the impact of adding audio information as additional input to CAG and to CONCAT-0. For all methods except for VAST–which takes raw audio input–we transcribe the audio of each video using whisper-v3 large (Radford et al., 2022). For the CONCAT-0 baseline, we provide the transcription as additional input alongside the instruction and frames to the VideoLLM. For CAG, we provide the transcriptions for each video together with the per-video summaries as input to the text-only aggregator LLM.

Table 16 presents the results. Similar to the previous experiment, we find that CONCAT-0 continues to yield poor quality articles, even with audio information. Notably, however, both CONCAT-0 and CAG consistently obtain worse results with audio inputs than without (Table 2)—despite the sizable fraction of audio-support claims in WIKIVIDEO. For CONCAT-0, this may partly be explained by the fact that the pretraining data for the VideoLLMs we study does not include audio transcripts, and thus the prompts that incorporate them are out-of-distribution. For CAG, we found that including audio transcripts tended to result in substantially shorter final articles (avg. $\sim$ 164 tokens) than omitting them (avg. $\sim$ 206 tokens)—suggesting that the former may be less thorough in their coverage of the event relative to the references, leading to lower metric scores (Appendix E has quantitative examples). Identifying ways to more effectively incorporate audio into the WIKIVIDEO task thus constitutes an intriguing direction for future work.

G Additional Results

We report more statistics in Table 17 to show the varying performance across model sizes and variations. We report LLaVA-Video-7B,72B and QwenVL2.5-3B,7B,72B as well as Qwen2.5-32B,72B for article synthesis.

Per-event type results In Table 18, we breakdown the argument F1 scores for each VideoLLM+CAG combination. Here we see the highest F1 scores in the most commonly recognizable events: elections and sports. These events are often professionally broadcast and the entities that participate in these events are “high-resource” visual concepts. However, in events like disasters and demonstrations, we see a decrease in F1, especially in exact match as there are no longer high-resource entities to identify or heavily populated OCR content.

H Human Analysis

To provide an upper-bound on model performance, we recruit 3 fluent english speakers to write 3 ar-

Video	Audio
On April 15, 2019, a catastrophic fire broke out at Notre-Dame Cathedral in Paris, France, causing extensive damage to the historic landmark. The fire, which began in the attic of the cathedral, rapidly spread, engulfing the roof and the iconic spire, which collapsed in a dramatic fall witnessed by a shocked crowd gathered outside. The cathedral, a masterpiece of Gothic architecture dating back to the 12th century, suffered significant structural damage, with the roof and spire completely destroyed. Despite the intense blaze, the main stone structure, including the two bell towers, remained standing. Firefighters worked tirelessly throughout the night to contain the fire, using advanced equipment and techniques to prevent further damage. The event drew global attention, with people around the world expressing sorrow and solidarity. In the aftermath, French President Emmanuel Macron pledged to rebuild the cathedral, and a fundraising campaign was launched to support the restoration efforts. The fire raised questions about the safety and preservation of historic landmarks and highlighted the cultural and spiritual significance of Notre-Dame to both France and the international community.	On April 15, 2019, a catastrophic fire broke out at Notre-Dame Cathedral in Paris, France, causing significant damage to the historic landmark. The blaze, which started in the attic, rapidly spread, engulfing the roof and the iconic spire, which subsequently collapsed. Thick plumes of smoke and intense flames were visible from afar, drawing crowds of onlookers who watched in distress and solidarity. The fire, which lasted for several hours, was eventually brought under control by hundreds of firefighters. Despite the extensive damage, the main structure of the cathedral, including the two bell towers, survived. President Emmanuel Macron pledged to rebuild the cathedral, and international support for the restoration efforts was swift and substantial. The incident highlighted the cultural and historical significance of Notre-Dame, one of the world's most recognized and cherished Gothic cathedrals.

Table 12: QwenVL+CAG with and without audio. Notre-Dame Fire

Video	Audio
The 2022 Georgia Senate election was a closely contested race between incumbent Democrat Raphael Warnock and Republican challenger Herschel Walker. After the initial election on November 8, 2022, neither candidate secured the required 50% majority, leading to a runoff election on December 6, 2022. In the initial vote, Warnock received 49.2% (1,944,893 votes) and Walker received 48.7% (1,927,402 votes), with a difference of 17,491 votes. An independent candidate, Chase Oliver, received 2.1% of the vote. The runoff election, which determined the ultimate winner, saw a significant turnout, with over 2 million early votes cast, highlighting the high level of voter engagement. The outcome of the runoff was crucial for the balance of power in the U.S. Senate, as it would determine whether the Democrats maintained their slim majority or if the Republicans gained control. Ultimately, Raphael Warnock was declared the winner, securing his re-election to the Senate.	The 2022 Georgia Senate election was a closely contested race between incumbent Democratic Senator Raphael Warnock and Republican challenger Herschel Walker. With neither candidate reaching the 50% threshold required to avoid a runoff, the election will proceed to a runoff on December 6, 2022. Preliminary results showed Warnock leading with 49.2% of the vote (1,944,893 votes) and Walker trailing slightly with 48.7% (1,927,402 votes). Independent candidate Chase Oliver received 2.1% of the vote (81,130 votes). The outcome of the runoff is expected to have significant implications for control of the U.S. Senate and for voting rights legislation, as well as for the broader political landscape in Georgia and nationally.

Table 13: QwenVL+CAG with and without audio. Georgia Senate Election

Video	Audio
Hurricane Irma was a powerful and destructive Category 5 hurricane that caused catastrophic damage across the Caribbean and parts of the southeastern United States in September 2017. Originating as a tropical wave, Irma rapidly intensified into a Category 5 storm with maximum sustained winds of 180 miles per hour, becoming the most powerful hurricane ever recorded in the open Atlantic. The storm made landfall in the Caribbean, particularly devastating the islands of St. Martin, the British Virgin Islands, and Barbuda. It then moved northwest, causing extensive damage in the Bahamas and Cuba before making landfall in Southwest Florida on September 10. The hurricane brought torrential rain, storm surges, and high winds, resulting in widespread flooding, structural damage, and power outages. In Florida, cities such as Miami Beach and Tampa experienced severe flooding and wind damage. The storm's impact was felt across multiple states, leading to widespread evacuations and a significant humanitarian response. Recovery efforts were complicated by the extensive damage to infrastructure, including roads, bridges, and power lines. The hurricane's aftermath highlighted the resilience of affected communities and the critical need for coordinated disaster preparedness and response.	Hurricane Irma was a powerful and destructive Category 5 hurricane that caused widespread damage across the Caribbean and the southeastern United States in September 2017. Originating as a tropical wave, Irma rapidly intensified into a Category 5 storm with maximum sustained winds reaching 180 miles per hour. The hurricane made landfall in the Leeward Islands, causing catastrophic damage, particularly on the islands of Barbuda and St. Martin. It then moved through the Turks and Caicos, the Bahamas, and eventually struck the Florida Keys and the southwestern coast of Florida. The storm brought torrential rainfall, storm surges, and strong winds, leading to extensive flooding and structural damage. In the aftermath, communities faced significant challenges, including the loss of electricity, water, and communication services. Recovery efforts were hampered by the scale of the destruction, and many areas required substantial aid and support to rebuild.

Table 14: QwenVL+CAG with and without audio. Hurricane Irma

Topic	**LLM**	**CAG**
Liberation Day Tariffs	Liberation Day Tariffs, also known as Vrijheidsdag Tariffs, are special rates and discounts offered by various businesses and organizations in the Netherlands on May 5th, Liberation Day (Vrijheidsdag). This day commemorates the end of the Nazi occupation during World War II in 1945. The tariffs typically include reduced prices on public transport, free admission to museums, and special events and activities designed to encourage participation and celebration among the Dutch population. These offers aim to make the commemoration accessible and inclusive, fostering a sense of community and national pride.	On April 2, 2025, President Donald Trump delivered a speech in Washington, D.C., announcing new tariffs on imports from several countries, including China, the European Union, Vietnam, Taiwan, Japan, India, South Korea, Thailand, Switzerland, Indonesia, and Malaysia. The tariffs, ranging from 20% to 36%, were designed to address what the administration perceived as unfair trade practices and to encourage domestic manufacturing. Trump emphasized that the U.S. would not impose a full 100% reciprocal tariff, stating his intention to be k̈ind to the worldẅhile protecting American industries. The announcement, made after the stock market closing bell, had already contributed to market volatility, with significant losses reported in the stock market, including a $3 trillion drop in value, referred to as a s̈tock market armageddon.F̈inancial analysts, such as those from JPMorgan Chase & Co., raised the recession risk forecast to 60%, up from 40% prior to the announcement. The legal and economic implications of these tariffs were widely discussed, with experts debating the constitutionality of the President's actions under the International Emergency Economic Powers Act (IEEPA) and the potential long-term impacts on global trade dynamics and consumer prices.
Blue Ghost Mission 1	The Blue Ghost Mission 1 was the inaugural flight of the Blue Ghost spacecraft, developed by Astra, an American aerospace manufacturer. Launched from the Mid-Atlantic Regional Spaceport (MARS) on Wallops Island, Virginia, the mission aimed to demonstrate the capabilities of the reusable suborbital vehicle designed for research and technology development. The spacecraft, which can carry payloads up to 150 kilograms, successfully reached its intended altitude and completed a series of tests before safely returning to Earth. The mission marked a significant milestone in Astra's efforts to provide cost-effective and flexible access to space for scientific and commercial applications.	The Blue Ghost M1 lunar lander, developed by Firefly Aerospace, successfully landed on the Moon on February 24, 2025, as part of NASA's Artemis program. The mission, which aimed to advance lunar exploration and scientific research, was a significant milestone in space exploration. The spacecraft, equipped with advanced scientific instruments, was designed and assembled at Firefly Aerospace's industrial facility, where engineers and technicians worked meticulously to ensure its precision and reliability. The landing process was closely monitored from the company's control room, where a team of experts tracked the lander's descent using detailed telemetry data. The successful touchdown was met with jubilation among the team members, marking a testament to their dedication and innovation in space technology. The mission not only contributed to the ongoing exploration of the Moon but also highlighted the capabilities of private aerospace companies in collaboration with government space agencies.
Canadian Federal Election	The 2025 Canadian federal election was held on October 21, 2025, to elect members of the House of Commons of Canada, the lower house of the bicameral Parliament of Canada. This election determined the composition of the 45th Canadian Parliament and the next Prime Minister of Canada. The Liberal Party, led by incumbent Prime Minister Justin Trudeau, faced off against the Conservative Party, led by Pierre Poilievre, among other smaller parties. Key issues included economic management, healthcare reform, and climate change policies. The election saw high voter turnout and significant debate over the direction of national policy, reflecting deep divisions within the electorate. Results from the election were closely watched, with implications for both domestic and international affairs.	The 2025 Canadian federal election, held on April 29, 2025, resulted in a fragmented Parliament with no single party achieving a majority. The Liberal Party, led by the incumbent Prime Minister, secured 119 seats, while the Conservative Party, under the leadership of Pierre Poilievre, gained over 20 seats, bringing their total to 83 and receiving the highest share of votes since 1988. The Bloc Qu0̆0e9b0̆0e9cois won 15 seats, the New Democratic Party (NDP) secured only 4 seats, and the Green Party failed to win any seats. The election saw significant shifts in voter support, particularly from the NDP to the Liberals, and created a challenging political environment with neither the NDP nor the Liberals able to form a coalition government. Pierre Poilievre, in his post-election speech, acknowledged the party's progress but emphasized the need for continued effort and change, looking ahead to future elections. The overall atmosphere during election night was one of excitement and celebration, especially among Liberal supporters, as they awaited the final results.

Table 15: Qualitative outputs on a sample of the 2025 subset of WikiVideo for the LLM and CAG outputs.

1. "Describe the video in detail and focus on the specific examples of CRISPR applications in medicine and agriculture mentioned, as well as the ethical considerations discussed."

2. "Describe the video in detail and focus on the specific demands of the protesters and any notable incidents or interactions during the convoy."

3. "Describe the video in detail and focus on the specific locations affected, the extent of damage caused, and any unique geological features observed during the eruption."

4. "Describe the video in detail and focus on the eruption's causes and effects in relation to the earthquake and tsunami."

5. "Describe the video in detail and focus on the specific mission details, such as the mission name, duration, objectives, and any unique features of the spacecraft or crew."

Figure 11: RePrompts from R1 Provided to a VideoLLM

You are an experienced Wikipedia editor. You will be shown summaries of one or more videos related to the same event. Your task is to write the lead section of a Wikipedia article based ONLY on the information provided in the video summary or summaries. The lead section MUST match the quality, style, and tone of real Wikipedia articles. DO NOT write in the style of a news journalist. DO NOT use any external sources or additional knowledge you have about the event. DO NOT output anything other than the Wikipedia lead section. DO NOT refer to any of the videos explicitly in your output. DO NOT write anything except for the Wikipedia lead section, even if the summaries are cut off. You MUST start your output with "<lead>". ONLY START YOUR REPORT WITH <lead>. DO NOT WRITE ANYTHING EXCEPT FOR THE WIKIPEDIA ARTICLE.

Figure 12: Prompt For Article Synthesis.

Method	VLM	R1	BS	Arg	AS
CAT-0	QVL	11.1	81.6	22.1	7.50
CAG-R	QVL	32.1	85.7	26.3	12.6
CAG-R	QVL	**34.0**	**86.4**	**30.8**	**14.3**

Table 16: WIKIVIDEO article generation results for CAG *with* audio inputs. Bottom row shows CAG *without* audio (copied from Table 2), which performs best.

ticles. These annotators receive the information request and the set of "oracle" relevant videos and are instructed to write the article from this information. This human annotation is fundamentally different than our data collection process because instead of grounding and 'discriminating' against an existing text, the annotators perform the same task as CAG taking the videos and writing information from them. To create the human generated articles, we provide annotators the relevant videos and instruct them to write the lead of a Wikipedia article. An interesting note from this experiment is we notice the annotators perform article writing similar to CAG, taking notes on each video before aggregating them in an article.

In Table 19, we baseline human performance against the original Wikipedia article as the predicted article and the best method (CAG +QwenVL). We observe that the current metrics for the task don't accurately capture the quality of the human written article, which has no hallucinations and is fully follows the constraint of only including video content. We show these results qualitatively in Table 20, Table 21, and Table 22.

Method	VideoLLM	R1	R2	RL	BS	Arg	AS
CONCAT-0	LLaVA-Video-7B	4.24	1.02	2.93	79.35	20.07	6.26
	LLaVA-Video-72B	7.34	1.60	4.78	71.99	19.31	5.08
	VAST	16.62	1.71	11.19	80.55	8.04	7.13
	InternVideo2.5	11.85	2.32	7.90	80.78	18.33	9.53
	QwenVL2.5-3B	9.60	2.36	6.27	80.80	20.22	7.73
	QwenVL2.5-7B	9.82	2.62	6.25	81.27	21.28	9.04
	QwenVL2.5-72B	11.34	3.13	7.06	81.60	23.72	8.01
CONCAT-REPROMPT	LLaVA-Video	6.36	1.51	4.22	80.03	21.34	5.50
	InternVideo2.5	6.93	1.68	4.83	79.62	22.48	6.19
	QwenVL2.5	8.38	2.71	5.49	81.94	22.89	7.17
CAG-0 +32B	LLaVA-Video-7B	31.04	7.96	18.14	85.65	24.49	18.71
	LLaVA-Video-72B	28.55	7.10	16.00	77.00	22.33	15.54
	VAST	16.80	1.12	11.18	81.63	9.17	14.01
	InternVideo2.5	28.11	6.06	16.94	84.80	22.21	16.59
	QwenVL2.5-3B	30.78	7.87	18.19	85.39	24.32	16.37
	QwenVL2.5-7B	31.64	7.88	18.13	85.59	24.35	16.31
	QwenVL2.5-72B	32.59	8.86	18.89	85.81	26.60	15.71
CAG-0 +72B	LLaVA-Video-7B	34.87	10.72	20.18	86.44	28.09	16.34
	LLaVA-Video-72B	30.02	8.68	17.59	77.59	26.21	13.51
	VAST	19.55	1.45	12.40	82.21	11.23	10.87
	InternVideo2.5	32.54	8.98	19.47	85.82	25.65	17.58
	QwenVL2.5-3B	32.92	10.00	19.37	85.95	27.44	16.27
	QwenVL2.5-7B	34.01	10.05	19.47	86.24	26.97	16.72
	QwenVL2.5-72B	33.58	10.15	19.15	86.18	28.97	15.63
CAG	LLaVA-Video	33.38	10.05	19.44	84.55	28.26	15.23
	InternVideo2.5	33.91	9.58	20.07	86.13	27.01	14.23
	QwenVL 2.5	33.96	10.90	19.45	86.35	30.77	14.29

Table 17: Vision only results by method.RePrompt results are **only** the outputs from the follow-up questions and does not include the generic captions.

	All		Sport		Disaster		Election		Social		Demonst		Discover		Political	
Model	ED	EX	ED	EX	ED	EX	ED	EX	ED	EX	ED	EX	ED	EX	ED	EX
LV	28	09	35	11	24	04	58	44	36	08	17	00	34	13	26	08
IV	27	06	40	17	24	03	49	25	42	00	13	00	30	09	23	04
QVL	31	11	50	32	29	06	62	50	38	08	18	00	31	10	25	05

Table 18: Argument F1 by MultiVENT-G Event Type for CAG Sport, Disaster, Election Social: Social Event, Demonst: Demonstration, Discover: Discovery OR Launch, Political: Political Development. ED: Edit Distance, EX: Exact Match

Method	VideoLLM	R1	BS	Arg	AS
CAG	QwenVL	40.57	86.77	30.80	14.29
RAG+CAG	QwenVL	23.84	77.85	20.65	9.01
Human Annotator		38.54	86.34	39.07	17.17
Wikipedia		64.53	90.47	62.82	23.53

Table 19: Comparison to human performance.

CAG	Human	Reference
The 2022 Georgia Senate election was a closely contested race between incumbent Democrat Raphael Warnock and Republican challenger Herschel Walker. After the initial election on November 8, 2022, neither candidate secured the required 50% majority, leading to a runoff election on December 6, 2022. In the initial vote, Warnock received 49.2% (1,944,893 votes) and Walker received 48.7% (1,927,402 votes), with a difference of 17,491 votes. An independent candidate, Chase Oliver, received 2.1% of the vote. The runoff election, which determined the ultimate winner, saw a significant turnout, with over 2 million early votes cast, highlighting the high level of voter engagement. The outcome of the runoff was crucial for the balance of power in the U.S. Senate, as it would determine whether the Democrats maintained their slim majority or if the Republicans gained control. Ultimately, Raphael Warnock was declared the winner, securing his re-election to the Senate.	On December 6th, 2022, ABC news declared Raphael Warnock the winner of the Georgia Senate Race. Warnock, the incumbent Democrat, defeated his opponent, republican Hershel Walker, in a runoff election by a little more than 1% of the vote. Warnock had initially won election in 2020, when he, along with President Joe Biden and Senator John Ossoff, led a surprise sweep of the presidential and senate elections in the historically conservative state. A month prior to the runoff, Warnock also held a slight lead over Walker on election day. At that point, with 99% of the votes counted, Warnock led by 35,429 votes, 49.4% to 48.5%. Despite this, in Georgia a runoff is triggered if no candidate wins at least 50% of the vote in the general election. With Libertarian candidate Chase Oliver pulling roughly 2% of the vote, MSNBC and CNN both reported that neither Warnock nor Walker were able to reach this threshold in November. This result was generally expected, as polling averages from the weekend prior to the initial general election showed an extremely tight race, from anywhere between a 0.8% margin from MSNBC to just a 0.1% margin from FiveThirtyEight. Many battleground senate races were decided relatively early on Election day, with North Carolina, Ohio, and Florida all being called for Republicans, while Pennsylvania and New Hampshire were called for Democrats. However, after these initial results, both Georgia and Wisconsin, where Ron Johnson held a slight lead over Mandela Barnes, were too close to call, while Nevada and Arizona also took longer for a victor to be declared. Interestingly, in contrast to the Georgia senate race, the Georgia gubernatorial election was decided without a runoff, with Republican Brian Kemp winning re-election in his rematch with Democrat Stacey Abrams, indicating a large number of split-ticket voters. This came despite Abrams' long-term investments into an activist-driven campaign. In the lead up to the runoff election, Walker heavily emphasized Joe Biden's low approval ratings and the economy, while Warnock focused on Walker's lack of knowledge about issues and allegations of violence. The incumbent senator's strategy focused on turning out the democratic base in cities, while the challenger focused on more conservative rural areas; both campaigns targeted moderate educated voters in the suburbs around Atlanta. In addition, Georgia Democrats won a key court battle to allow for early voting in runoff elections. This decision challenged SB 202, a 2021 state law that limited early voting periods and restricted weekend voting.	The 2022 United States senate election in Georgia was held on November 8, 2022 to elect a member of the U.S. senate to represent the state of Georgia. The election was part of the 2022 midterm elections and featured incumbent democratic senator Raphael Warnock and Republican former football player Herschel Walker. Warnock, who had one a shortened seat in a special election held in 2020-21 defeated Walker in a runoff election and was re-elected for his first full term in office. Under Georgia's two-round election system, a runoff election occurs if no candidate receives more than 50% of the vote. In the November 8 election, Warnock received 49.4% of the vote and Walker received 48.5%, triggering the December 6th runoff election. Warnock's victory in the runoff was a state-wide victory for Democrats. Warnock's win gave Democrats a net gain of one seat in the 2022 midterms. The election was also one of many very close Senate elections during the 2022 midterms, with both Nevada and Wisconsin also having tight races.

Table 20: Georgia Senate Election

CAG	Human	Reference
A significant earthquake with a magnitude of 7.0 struck Anchorage, Alaska, causing widespread disruption and damage. The National Weather Service confirmed the event and reassured the public that there was no tsunami threat. Footage from the aftermath showed damaged roads, disrupted infrastructure, and scenes of chaos, including a supermarket where items had fallen off shelves. The earthquake prompted school evacuations and the temporary closure of the Anchorage airport, as the air traffic control tower was evacuated. While the full extent of the damage remains unclear, ongoing assessments and recovery efforts are underway to address the impact of the quake.	A magnitude 7.0 earthquake hit southcentral Alaska on Friday, November 30, 2018. The earthquake hit approximately 7 miles north of Anchorage at 8:29 AM. Anchorage is Alaska's largest city, with half of the state's population living in the region. Following the earthquake, the National Tsunami Warning Center issued and then canceled a tsunami warning for the coastal zones of southern Alaska. Governor Bill Walker issued a disaster declaration. In the three hours after the earthquake, the US Geological Survey recorded at least 30 aftershocks with magnitudes ranging from 2.7 to 5.7; hundreds of aftershocks were eventually recorded. During the earthquake and its aftershocks, students and office workers sheltered in place under their desks, and some buildings were evacuated, including the air traffic control tower at the Anchorage airport. The earthquake caused major infrastructure damage across the city, according to the Anchorage police department. Impacts include damage to roadways and water mains, visible cracks in buildings, stores and homes in disarray, and one house fire caused by a damaged gas pipe. No deaths have been reported.	On November 30, 2018 at 8:29 a.m. AKST (17:29 UTC), a magnitude 7.1 earthquake hit Anchorage in South Central Alaska. The earthquake's epicenter was 10 miles north of Anchorage and occurred at a depth of 29 miles. It was followed by a magnitude 5.7 aftershock. The National Tsunami Warning Center issued tsunami warnings for nearby costal areas, including Cook Inlet. The warnings were lifted shortly after being issued.

Table 21: Anchorage Earthquake

CAG	Human	Reference
Hurricane Irma was a powerful and destructive Category 5 hurricane that caused widespread damage across the Caribbean and the southeastern United States in September 2017. Originating as a tropical wave, Irma rapidly intensified into a Category 5 storm with maximum sustained winds reaching 180 miles per hour. The hurricane made landfall in the Leeward Islands, causing catastrophic damage, particularly on the islands of Barbuda and St. Martin. It then moved through the Turks and Caicos, the Bahamas, and eventually struck the Florida Keys and the southwestern coast of Florida. The storm brought torrential rainfall, storm surges, and strong winds, leading to extensive flooding and structural damage. In the aftermath, communities faced significant challenges, including the loss of electricity, water, and communication services. Recovery efforts were hampered by the scale of the destruction, and many areas required substantial aid and support to rebuild.	Hurricane Irma was a record-setting Hurricane that struck the south Atlantic region in early September, 2017, and that caused widespread damage. At the time, Irma was the most powerful Hurricane recorded in the open Atlantic region and was the second-strongest ever to hit Cuba. Beginning as a tropic wave, Irma struck many places—including the Lesser Antilles, Dominica, Guadeloupe, Barbuda, Antigua, St. Martin, Puerto Rico, Haiti, the Dominican Republic, and Cuba—as a Category 5 storm—as well as Florida as a Category 3 storm. At its peak intensity, Irma reached more than 180 mph winds and caused severe flooding and damage to buildings and infrastructure throughout the region.	Hurricane Irma was an extremely powerful Cape Verde hurricane that occurred in early September 2017. Irma was a Category 5 hurricane and the most powerful hurricane on record in the open Atlantic region outside of the Caribbean Sea and Gulf of Mexico. Irma was also the strongest tropical cyclone by wind speed worldwide in 2017. Irma caused widespread and catastrophic damage throughout its path, and was particularly severe in the northeastern Caribbean. Irma developed from a tropical wave near the Cape Verde Islands. Irma then became a Category 3 hurricane on the Saffir-Simpson wind scale before resuming intensifying on September 4 and becoming a Category 5 hurricane by early September 5. Irma's intensity peaked on September 6 with 1-minute sustained winds at 180mph and a minimum pressure of 914 hPa. Before making landfall in Cuba, Irma weakened to a Category 4 hurricane, but regained its Category 5 status before hitting Cuba. Irma hit both Caribbean islands and the continental United States. Irma caused catastrophic damage in Barbuda, Saint Barthélemy, Saint Martin, Anguilla, and the Virgin Islands as a Category 5 hurricane. Irma also made landfall in Anguilla, Barbados, Cuba, French West Indies, Haiti, Puerto Rico, and the Dutch side of Sint Maarten. After crossing the Straits of Florida, Irma made landfall in Cudjoe Key on September 10 making landfall in Florida before then making landfall at Marco Island.

Table 22: Hurricane Irma