MovieRecapsQA: A Multimodal Open-Ended Video Question-Answering Benchmark

Shaden Shaar* Bradon Thymes* Sirawut Chaixanien Claire Cardie
Bharath Hariharan
Cornell University
{ss2753, bmt63}@cornell.edu

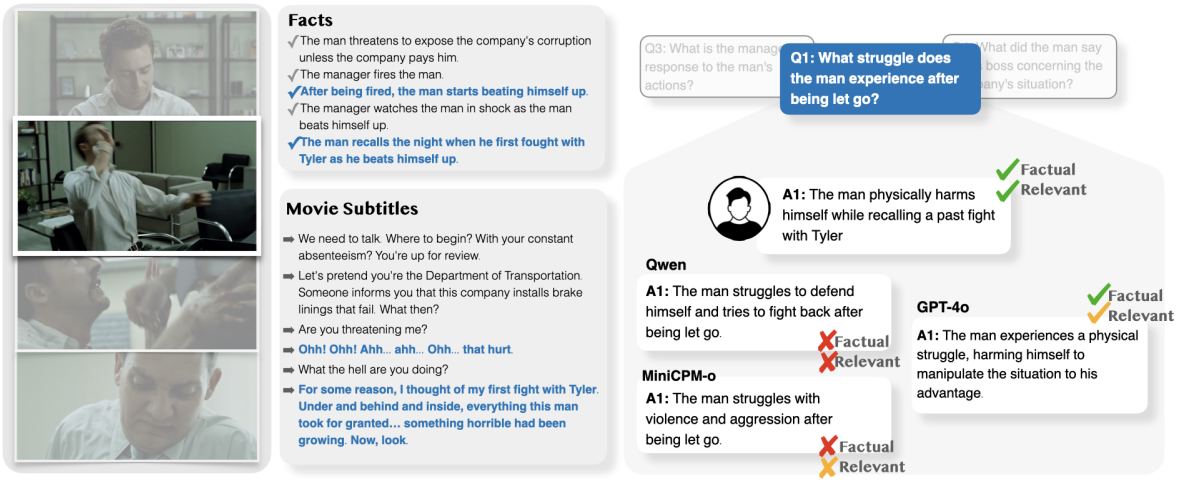

Figure 1. **MovieRecapsQA Benchmark.** An example (movie *Fight Club*) from our benchmark illustrating how MLLMs answer questions using recap-video frames and aligned movie subtitles. Q1 was constructed from Facts (3) and (5), and answering it requires integrating visual cues (Frame 2) with supporting subtitle evidence (Lines 4 and 6). We show the human answer alongside model outputs, evaluated for relevance and factuality on a 0–5 fact-grounded scale, with colors indicating quality from lowest to highest: X, X, ✓, ✓.

Abstract

Understanding real-world videos such as movies requires integrating visual and dialogue cues. Yet existing VideoQA benchmarks struggle to capture this multimodal reasoning and, given the difficulty of evaluating free-form answers, largely resort to simple multiple choice questions. We introduce a novel open-ended multimodal VideoQA benchmark, MovieRecapsQA, created using movie recap videos – a distinctive type of YouTube content that summarizes a film via a voiceover description of key clips from the movie (recap video). From the transcribed voiceover (recap summary) of 60 recap videos, we generate ≈8.2K questions along with the necessary "facts" expected in each answer; the former facilitates the creation of questions that require mutimodal reasoning and the latter allow the construction of a reference-free evaluation metric that can be applied to open-ended responses. To our knowledge, this is the first reference-free open-ended VideoQA benchmark. The benchmark allows each question to be evaluated in different input video settings: given (a) the full-length movie, (b) the full (≈11 min) recap video (visual only), (c) ≈14 min of aligned movie scenes, i.e, movie scenes relevant to the question, and (d) ≈1.2 min of aligned recap video scenes. In all cases, the text of any associated movie dialogue is provided. Each question is categorized by the modality required to answer it—visual, dialogue, or both—enabling fine-grained evaluation of multimodal capabilities. We benchmark (setting (d)) seven state-of-the-art MLLMs and find that (i) only our reference-free metric produces meaningful human-aligned model separation; (ii) vision-centric questions yield the lowest scores across all models; (iii) removing visual input often improves model factuality; and (iv) the primary bottleneck is visual perception, not visual reasoning.

*These authors contributed equally to this work

1. Introduction

Consider the prospect of a system that can watch a movie like *Fight Club* (Figure 1) and then answer questions about the movie. To answer a question like "What struggle does the man experience after being let go?", the system would need to listen to the dialogue to determine when the person is being fired (i.e., "let go"), and then look at the corresponding video sequence to see how the person reacts. The need for multimodal reasoning across video and dialogue is not limited to the context of movie understanding; it generalizes to real world tasks, such as robot reasoning over visual and linguistic modalities.

Most existing Video Question Answering (VideoQA) benchmarks fall short of capturing the complexity of such multimodal understanding. Many benchmarks focus on a single modality (e.g., solely visual or solely audio) [27] and often feature simple thematic questions and relatively short video clips. Even recent work that explicitly focuses on long videos and multimodal reasoning uses simpler multiple-choice questions [12, 33, 36]. In contrast to open-ended question answering, multiple choice questions offer shortcuts for answering the question without understanding the video. Thus, current benchmarks provide only limited understanding of how well models generalize to real-world content, such as movies and real-life social contexts, where understanding the temporal dynamics and fine-grained character interactions require complex multimodal reasoning.

A key reason for the lack of better benchmarks is the sheer difficulty of acquiring complex questions and answers — careful manual construction of high-quality QA pairs is time-consuming. Evaluating model answers in this open-ended setting is also hard: gold standard answers are not typically single words or simple phrases, and comparing them to model-generated answers is subjective. As a result, evaluation strategies rely on the use of word-based semantic similarity metrics (e.g., ROUGE) that have been shown to have a low correlation with human judgments on factuality and quality [9, 18, 30].

The evaluation challenge is further magnified in the video domain. In text-based question answering, using LLMs as a judge has been proposed to verify an answer's factuality and relevance with respect to a source text. However, this approach is not directly transferable to VideoQA: using the full video as "context" for a judge is computationally expensive and imprecise. The model would have to parse complex, multimodal signals just to find the relevant information for answer grounding. Consequently, existing attempts to use LLM judges for VideoQA often fall back on the reference-answer paradigm [5, 26].

We address these challenges by tapping into a novel data source: *movie recaps*. A movie recap is a distinctive type of YouTube content that provides a short (∼10 min) summary of a full-length film via a voiceover description of key visual-only clips from the movie. We will refer to the clips as the **RecapVideo** and the transcribed voiceover as the **RecapSummary**. Importantly, the RecapSummary can be automatically mined to extract facts, questions and answers. And because it is aligned with the RecapVideo, extracted questions and answers can be grounded in specific segments of the movie and its corresponding dialogue. We leverage this insight to introduce MovieRecapsQA, a novel benchmark that features open-ended questions requiring multimodal reasoning over long-form, narrative video. Multiple levels of granularity of input video length are provided for each QA pair ranging from the full-length movie to just the RecapVideo segments relevant to the question.

Crucially, MovieRecapsQA is designed to solve the core evaluation problems of open-ended questions in video. Specifically, we introduce an intermediary annotation layer of atomic facts—concise, verifiable statements—that are automatically derived from the RecapSummary and used to generate a QA pair. The facts, in turn, support our proposed a reference-free evaluation that uses an LLM judge grounded with question-specific facts to evaluate a generated answer. The facts provide a precise, text-based, verifiable representation of video content that enables the assessment of answer factuality, relevance, and coherence without relying on restrictive reference answers.

Using MovieRecapsQA, we benchmark seven state-of-the-art multimodal large language models (MLLMs), both proprietary and open-source, alongside human participants. The dataset exposes several trends: (i) semantic and reference-based metrics often rank models similarly with little separation, whereas our reference-free metric yields more meaningful separation and better alignment with human preferences; (ii) vision-centric questions yield the lowest scores across all models; (iii) proprietary model factuality *improves* when visual input is removed, revealing a visual understanding gap; and (iv) the primary bottleneck is visual perception, not reasoning, with fine-grained scene details showing the largest human–model gap.[1]

Our contributions can be summarized as:

1. We introduce **MovieRecapsQA**, a long-form, multimodal VideoQA benchmark comprising ≈8,200 open-ended questions across 60 videos with modality-type labels (dialogue-centric, vision-centric, multimodal) and reasoning categories.
2. We propose a **reference-free evaluation metric** for open-ended VideoQA that uses recap-derived atomic facts to ground an LLM judge in assessing factuality, relevance, and coherence without relying on a unique reference answer.
3. We benchmark and analyze the performance of 7 multimodal LLMs and human annotators on MovieRecapsQA, revealing a substantial human–model performance

[1] All code is released, MovieRecapsQA.

gap, a pronounced over-reliance on dialogue, and systematic failures on fine-grained visual perception.

2. Related Work

Video Question Answering (VideoQA). VideoQA was introduced as a proxy task to evaluate a model's ability to understand and reason over video inputs. Existing VideoQA benchmarks typically adopt either a multiple-choice format (e.g., TVQA [21], MovQA [40], How2QA [37], DramaQA [6]) or an open-ended format (Video-Bench [29], MVBench [23], EgoSchema [27], CinePile [33]). Open-ended benchmarks, however, remain both more challenging and far less common, as evaluation for free-form responses is still an open problem [2].

Moreover, there has been growing interest in multimodal VideoQA [33, 40], where models receive both video and textual context – and questions typically require integrating information across visual and textual modalities to generate accurate responses. However, curating such datasets is far more expensive, and they often need to be partially or fully automatically generated, with some generation methods limiting the resulting QA pairs to either visual-only or dialogue-only questions.

Open-Ended Question Answering Evaluation Evaluating the quality of open-ended, free-form text remains one of the most significant challenges in QA, and other generation tasks. For years, the field relied on n-gram overlap metrics, such as ROUGE[24], BLEU[31], METEOR[4]. Later embedding-based metrics such as BERTScore [41] and BARTScore were introduced, though they showed limited alignment with human judgments [9]. More recently, open-ended QA benchmarks have exclusively used LLM-as-a-judge metrics (e.g., HELMET [38] and GEval [25]) to evaluate the relevance and coherence of model answers. Both approaches, however, require reference answers. To our knowledge, there are no benchmarks that evaluate model answers without relying on a reference, a limitation that prevents the creation of larger and more cost-effective datasets.

In other QA settings (e.g., text-based QA), the factuality of an answer is a critical evaluation dimension but remains difficult to measure reliably [8, 20]. Early approaches relied on NLI models or entity matching [11], yet these methods consistently failed to capture factuality with strong human alignment [19]. More recently, LLM-as-judge factuality metrics (i.e., FactScore[28] and VeriScore[34]) assess factuality without requiring a reference answer more reliably. Since factuality concerns the truthfulness of the answer—not its similarity to a reference—these methods instead verify claims against the input text context, with some work suggesting that context quality affects scoring[32]. This is particularly difficult to achieve in VideoQA, as using raw video as the verification context is unreliable (with MLLM) and computationally expensive. To our knowledge, no VideoQA dataset explicitly measures factuality.

3. MovieRecapsQA Benchmark

We construct our dataset automatically by leveraging a widely available and increasingly popular genre of YouTube content known as recap videos. These videos narrate the full storyline of a movie—typically in an 8–15 minute continuous format—while replaying key scenes. Unlike Wikipedia Synopsis or IMDb plot summaries, which are often high-level and omit substantial narrative detail, recap videos provide dense, scene-by-scene coverage of the film's major events, characters, and plot developments.

Additionally, unlike other VideoQA datasets built from movies (e.g., MovieQA, TVQA, MovQA), the recap videos directly pair each narrated event with the corresponding movie shot(s) through visual-only representation (i.e., no audio). This tight coupling between narration and visuals allows for accurate summary–movie alignment and enables a more accurate pairing between the question, the movie dialogue, and the question-specific facts derived from the RecapSummary. The dataset was constructed in two main steps: (1) Collection & Alignment, and (2) QA Generation.[2]

3.1. MovieRecapsQA Collection & Alignment

We begin by selecting a set of 60 films released between 1980 and 2024, spanning both widely known titles (e.g., Avatar) and more niche works (e.g., Year One).[3] To source recap content, we compile a list of the top 10 most popular YouTube recap channels, which we treat as trusted providers of high-quality summaries. Using the YouTube API, we retrieve candidate RecapVideos for each movie and collect the top five search results originating from these trusted channels. Finally, we manually filter these videos, verifying that each recap is indeed aligned with the corresponding movie's storyline and title.

Next produce a scene-level alignment of each RecapVideo to its corresponding movie. We first apply SceneDetect [16] to segment both the full movie and its corresponding RecapVideo into scenes. We group all consecutive frames belonging to the same scene. Then we embed the first and last three seconds of footage for every detected scene in both the movie and the RecapVideo, and compute cosine similarity to match corresponding shots. We use SlowFast [10] for the embeddings. Because some RecapVideos are not strictly chronological, we additionally perform a lightweight statistical alignment step to enforce a semi-chronological ordering of matched scenes. Further alignment details and results can be found in Appendix C.

[2] We address all copyright issues in Appendix A.

[3] We intentionally include movies with publicly available scripts so that future work can incorporate script-based annotations and build further extensions on top of our benchmark.

Dataset	Annotation	# QAPairs	Avg. Len. (s)	Modality	Q Type
MovieQA [35]	Human	6,462	203	No	MC
TGIF-QA [17]	Auto/Human	165,165	3	No	MC & OE
TVQA [21]	Human	152,545	76	No	MC
DramaQA [6]	Human	17,983	91	No	MC
MoVQA [40]	Human	21,953	992	No	MC
CinePile [33]	Auto	303,828	160	No	MC
MovieRecapsQA (Ours)	Auto	8,231	660	Yes	FF

Table 1. Comparison of existing video QA datasets. Our proposed dataset introduces multimodal distinctions and free-form question types, distinguishing it from prior benchmarks.

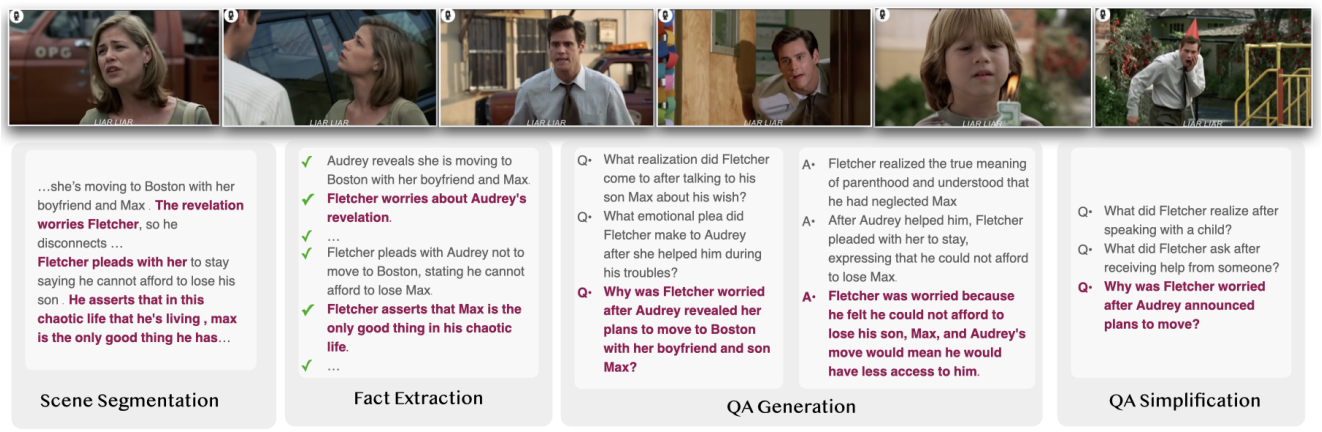

Figure 2. **QA Generation Pipeline.** Example question-answer pairs from recap video 6Tfmy3uGTmQ (for *Liar Liar*) on the recap segment "00:06:50–00:12:42" (and "00:31:44–00:50:12" from the movie). The red-highlighted text indicates the recap-segment input used to extract facts and generate the corresponding QA pair.

Statistic	Value
Dataset Size	
Total Films & Recap Videos	60
Total QA Pairs	**8,231**
Total Aligned Atomic Facts	**16,462**
Temporal Statistics	
Avg. Length of Full Recap Video	~660 s (11 min)
Avg. Length of Full Movie	~6,446 s (107 min)
Avg. Segment Length (Recap)	~73 s
Avg. Aligned Segment (Full)	~863 s (14 min)
Facts and Questions	
Avg. Total Facts from Segments	~12
Avg. Facts per Video	~222
Avg. Questions from Segments	~172

Table 2. Overall statistics for the MovieRecapsQA dataset.

Using this process, we obtain an alignment of the RecapVideo not only with the movie video but also with the corresponding dialogue via the available movie subtitles. Because the RecapVideo is itself aligned with the RecapSummary, we also obtain an alignment of the summary with the movie and its dialogue.

3.2. MovieRecapsQA QA Generation

To construct the question–answer pairs, we divide each RecapSummary into segments and use prompting to extract all facts from each recap segment via LLM-based prompting. We then instruct an LLM to generate QA pairs that rely on one or more facts within a single segment. However, because these extracted facts tend to be highly verbose, the resulting QA pairs are often too easy: they reveal excessive information about the scene. In addition, questions frequently include character names, which can hinder alignment with subtitles (since dialogue excerpts may not mention the character's name at the corresponding moment). To address these issues, we additionaly generated simplified QA pairs using a dedicated LLM prompt designed to abstract away identifying details. All dataset construction steps are performed using GPT-4.1. All prompts regarding the QA generation can be found in Appendix B.

Figure 2 illustrates QA generation for Segment 4 of the Liar Liar RecapVideo: fact extraction produces 32 facts (only four are shown); then verbose QA pairs and their simplified counterparts are generated. We retain the answer from the verbose QA pair and the question from the simplified QA pair in our final QA set.

3.3. Statistics

The MovieRecapsQA dataset is built on 60 films, resulting in a corpus of 8,231 open-ended question-answer pairs. Each question is aligned with with a specific segment from a movie recap video, the corresponding (video-only) segment from the original full-length movie, and the corresponding movie subtitles. A summary of the dataset's key statistics is presented in Table 3.

Questions are categorized by the primary modality required to answer them and the type of reasoning involved. The distribution of question modalitites is diverse, as shown in Table 3, which forces the models to use different types of reasoning. The breakdown of different question categories, based on the CinePile taxonomy, is also detailed in Table 3. All prompts used to categories the dataset can be found in Appendix E.

Question Modality	
Dialogue-centric	2,932
Multimodal	3,525
Vision-centric	1,774
Question Categories	
Narrative and Plot Analysis (NPA)	3,294
Character and Relationship Dynamics (CRD)	3,149
Thematic Exploration (TH)	677
Setting and Technical Analysis (STA)	655
Temporal Reasoning (TEMP)	456
Total	**8,231**

Table 3. Question statistics for MovieRecapsQA: modality distribution (top) and categories (bottom) over 8,231 total questions.

4. Evaluation Metrics

Evaluating free-form open-ended VideoQA is fundamentally harder than evaluating multiple-choice or short-answer formats. A correct open-ended answer must satisfy multiple criteria simultaneously: it must be factually accurate with respect to the video content, relevant to the question being asked, and internally coherent. No single existing metric captures all three dimensions, and most fail to capture even one reliably. We first survey the shortcomings of standard metrics, then introduce our reference-free evaluation framework designed to address them.

4.1. Baseline Metrics

Semantic Evaluation Metrics We first establish baselines using standard semantic similarity metrics that compare model-generated responses against reference answers at the lexical and embedding level. **ROUGE** [24] measures n-gram overlap; **BERTScore** [41] computes embedding-based similarity; and **BARTScore** [39] uses a seq2seq model to score conditional likelihood. These metrics capture surface form and paraphrastic similarity but not factual correctness.

Reference-Based Evaluation Metrics To assess factual correctness in long-form answers we employ reference-based LLM judges that leverage the reasoning capabilities of LLMs to evaluate response quality.

We include two reference-based LLM judges: **G-Eval** [25], a framework that rates the coherence and consistency of an answer given a reference, and **HELMET**[38], which targets two dimensions — *Fluency*, which captures grammatical correctness and coherence, and *Correctness*, which is designed to measure factual agreement with the reference answer on a 0–3 scale. HELMET Correctness serves as our primary reference-based baseline as it explicitly targets factual accuracy. However, it still relies on a single gold standard answer and cannot verify whether a model's response is grounded in the underlying video content.

4.2. Reference-Free Evaluation Metric

The limitations above motivate our central evaluation contribution: a reference-free LLM judge that evaluates whether a response is factually grounded in the video content, independent of how any particular reference answer is formulated.

We first leverage our recap pipeline to construct a *textual layer of atomic facts* that serves as a compact, verifiable representation of the video content.

For each question q, we collect a set of *atomic facts* $\mathcal{F}_q = \{f_1, \ldots, f_K\}$ derived from the aligned recap summary for the corresponding video segment. Each atomic fact is a short, standalone proposition about the movie (e.g., "*Tyler threatens to destroy the narrator's apartment*"), written so that its truth can be directly verified from the recap. In addition to the atomic facts extracted for each questions, we also extract a set of *claims* for each model response $C_r = \{c_1, \ldots, c_K\}$. Given a question q, the model claims C_r, the associated atomic facts $\mathcal{F}_q$, and the subtitles for the segment our LLM judge is prompted to evaluate a along three dimensions:

- **Factuality** (s_{fact}): to what extent the claims in a are supported (or contradicted) by $\mathcal{F}_q$.
- **Relevance** (s_{rel}): whether a directly addresses q and avoids introducing unsupported, off-topic content.

All reference-free evaluations are performed using GPT-4.1 mini as the LLM judge. We select this model for its cost efficiency given the scale of evaluation, as scoring 8.2K questions across seven models and three dimensions requires a substantial number of API calls. We verified that GPT-4.1 mini produces scores comparable to GPT-4.1 on our evaluation dimensions.

The judge returns integer scores in the range 0–5 for each dimension, where higher scores indicate better performance. Full prompts can be found in Appendix F.

We exclude the coherence metric from the main results, as the answers are too short to meaningfully exhibit internal contradictions. Results on the coherence dimension are reported in Appendix G.

5. Experiments

Our goal is to evaluate how current multimodal LLMs handle long-form, open-ended VideoQA under controlled, comparable conditions. This section describes how we construct model inputs, how we prompt each system, and how we run our ablations.

5.1. Models

We evaluate a diverse set of multimodal models, spanning both proprietary and open-source systems. Our selection includes leading proprietary models: GPT-4o[14], Gemini 2.5 Flash[7], Claude 3.5 Sonnet[1], and Amazon Nova Lite[15], which represent the current frontier in commercial video understanding capabilities. We complement these with open-source alternatives: LLaVA NeXT-Video[22], MiniCPM-o[13], and Qwen 2.5-VL[3]. All models are evaluated in a zero-shot setting with a standardized prompting scheme and input format.

5.2. Input Formats and Modalities

Each question in MovieRecapsQA is aligned with a specific recap segment, its corresponding full-movie segment, and the matched dialogue (subtitles + script) for that interval (Section 3). Unless otherwise noted, all models are evaluated using the recap segment as the visual source.[4]

Visual input. For models that accept raw video files (e.g., Gemini 2.5 Flash, Amazon Nova Lite), we pass the aligned recap segment as a short MP4 clip. For frame-based models (e.g., GPT-4o, Claude 3.5 Sonnet, LLaVA-NeXT-Video, MiniCPM-o, Qwen2.5-VL), we uniformly sample frames from the recap segment and pass them as an ordered image sequence. We use a different number of frames for each model, up to the maximum context length allowed by that model.

Dialogue input. We use the aligned subtitles from the same temporal window as the visual input.

Multimodal, frames-only, and dialogue-only conditions. By default, models in the *multimodal* setting receive both the visual input (video or frames) and the aligned dialogue. For our ablations, we also evaluate two restricted conditions: (i) *frames-only*, where the model receives only the visual input and no text, and (ii) *dialogue-only*, where the model receives only the subtitle snippet without any images or video. These three conditions allow us to tease apart how much each system relies on vision versus language, and whether multimodal input improves or harms factual accuracy.

[4]The dataset design enables multiple task formulations, with variable video length (clip, recap, full movie) and variable text length (segment-level or full dialogue). Further details can be found in Appendix D.

5.3. Human Study

To establish a human upper bound and to validate our evaluation metrics, we conducted a human study. We randomly sampled 118 questions from our dataset, spanning all three modalities (dialogue-centric, vision-centric, multimodal) These questions were presented to five human participants, who provided open-ended answers under the same conditions as models: each participant saw the question, the recap segment, and the aligned subtitles for that segment.

We then scored human answers using HELMET Correctness and our fact-based Factuality and Relevance metrics. For each question, we compute both the *average* human score (averaged across participants) and the *best* human score (maximum across participants).

6. Results & Discussion

This section presents an analysis of model performance on the MovieRecapsQA benchmark. Table 4 reports overall performance across all metrics and models, and Table 5 presents ablation results across question categories and reasoning types; further ablation details are provided in Appendix H.

6.1. The Metric Divergence Problem

Observation #1: Semantic scores fail to discriminate between models. Semantic metrics show virtually no discrimination across models; for example, ROUGE-L ranged only 0.22–0.28, BERTScore 0.63–0.69, and BARTScore 0.03–0.05, as shown in Table 4. Moreover, the per-model variance of each metric is ≤ 0.03, indicating that these metrics assign nearly identical scores to every answer produced by the same model, regardless of question or answer content. Together, this suggests that n-gram overlap and embedding-based similarity fail to capture anything beyond surface-level token matching.

Observation #2: Reference-based LLM judges do not align with human preference. While reference-based metrics yield better model separation than semantic metrics, they do not align with human preference, Table 4. For example, HELMET-Correctness rates MiniCPM-o, 1.27, above the best, 1.26, and average human, 0.98. Similarly LLaVA-NeXT-Video, 0.98, is scored on par with the average human. This suggests that reference-based judges inflate model scores and fail to reflect the true gap between model and human performance.

Observation #3: Reference-free metric produces human-aligned model separation. In contrast to both prior metric families, our reference-free metrics produce a wider and more meaningful spread, Table 4. Factuality scores range

	Semantic Metrics			Reference-Based Evaluation			Reference-Free Evaluation	
Model	**ROUGE-L**	**BERTScore**	**BARTScore**	**G-Eval**	**HELMET Fluency**	**HELMET Correctness**	**Factuality**	**Relevance**
LLaVA-NeXT-Video	$0.23_{\pm 0.01}$	$0.65_{\pm 0.01}$	$0.03_{\pm 0.00}$	$0.26_{\pm 0.03}$	$0.96_{\pm 0.04}$	$0.98_{\pm 0.89}$	$2.96_{\pm 1.97}$	$3.35_{\pm 1.47}$
Mini-CPM-o	$0.24_{\pm 0.02}$	$0.65_{\pm 0.01}$	$0.04_{\pm 0.00}$	$0.30_{\pm 0.04}$	$0.94_{\pm 0.05}$	$1.27_{\pm 1.08}$	$3.21_{\pm 2.04}$	$3.61_{\pm 1.47}$
Qwen2.5VL	$0.26_{\pm 0.02}$	$0.67_{\pm 0.01}$	$0.04_{\pm 0.00}$	$0.31_{\pm 0.03}$	$0.97_{\pm 0.03}$	$1.23_{\pm 0.98}$	$3.47_{\pm 1.98}$	$3.83_{\pm 1.41}$
Amazon Nova Lite	$0.28_{\pm 0.02}$	$0.69_{\pm 0.01}$	$0.05_{\pm 0.00}$	$0.32_{\pm 0.04}$	$0.99_{\pm 0.01}$	$1.29_{\pm 1.03}$	$3.53_{\pm 1.96}$	$3.93_{\pm 1.35}$
Claude 3.5 Sonnet	$0.22_{\pm 0.02}$	$0.63_{\pm 0.01}$	$0.05_{\pm 0.00}$	$0.37_{\pm 0.05}$	$0.98_{\pm 0.02}$	$1.35_{\pm 1.42}$	$3.76_{\pm 1.80}$	$3.92_{\pm 1.19}$
Gemini-2.5-Flash	$0.22_{\pm 0.02}$	$0.63_{\pm 0.01}$	$0.05_{\pm 0.00}$	$0.38_{\pm 0.05}$	$0.95_{\pm 0.05}$	$1.82_{\pm 1.33}$	$3.26_{\pm 2.35}$	$3.70_{\pm 1.57}$
GPT-4o	$0.28_{\pm 0.03}$	$0.68_{\pm 0.01}$	$0.05_{\pm 0.01}$	$0.37_{\pm 0.06}$	$0.94_{\pm 0.05}$	$1.43_{\pm 1.18}$	$3.99_{\pm 2.01}$	$3.97_{\pm 1.92}$
Avg. Human*	$0.16_{\pm 0.01}$	$0.88_{\pm 0.00}$	–	–	$0.94_{\pm 0.06}$	$0.98_{\pm 1.06}$	$4.01_{\pm 1.70}$	$4.01_{\pm 1.34}$
Best Human*	$0.19_{\pm 0.01}$	$0.87_{\pm 0.00}$	–	–	$0.93_{\pm 0.06}$	$1.26_{\pm 1.21}$	$4.59_{\pm 0.63}$	$4.53_{\pm 0.76}$

Table 4. **Models Performance.** This table reports model performance on our benchmark across semantic, reference-based, and reference-free metrics. We additionally include HUMAN* performance on a sampled set of 118 questions. For each metric, we report the mean score across all benchmark questions ± variance. HELMET scores range from 0–3, our reference-free metrics from 0–5, and all remaining metrics are normalized to 1.

from 2.96–3.99 across models, complimented with 4.01 and 4.59 for average and best human. Furthermore, the higher per-model variance of our metric (1.97–2.35 on FACTUALITY) relative to reference-based metrics (0.89–1.42) confirms that our scores reflect genuine differences in answer quality across questions rather than model-level tendencies.

6.2. The Modality Performance Gap

Observation #4: Vision-centric questions yield the lowest Factuality scores across all models. Questions requiring visual understanding, whether *vision-centric* or *multimodal*, score lowest on FACTUALITY for both models and humans, Table 5. Proprietary models drop from 3.63 (dialogue) to 3.15 (vision); open-source models from 3.21 to 3.05. Humans follow the same trend, falling from 4.17 to 3.84. This confirms that MOVIERECAPSQA poses a genuine visual challenge: getting the facts right is hard not because models are weak, but because visual information is difficult to extract even for humans watching the same video.

Observation #5: Models stay on-topic but fail to get the facts right. Unlike Factuality, Relevance scores remain stable across modalities for proprietary, 3.84–3.63, and open-source, 3.52–3.61, indicating that models consistently attend to the correct scene regardless of modality. Yet humans score substantially higher on Relevance across modality (≈4.0+), revealing a clear dissociation: models discuss the right content but cannot extract the precise visual facts needed to answer correctly, i.e., they know *where* to look, but not *what* to say.

6.3. The Visual Information Gap

Observation #6: Removing visual input improves proprietary model performance. Proprietary models score higher under dialogue-only than full-modality input across all question types and both metrics (Table 5). FACTUALITY improves by +0.49, +0.73, and +0.70 on dialogue, vision-centric, and multimodal questions respectively, with RELEVANCE following the same trend. The gains are largest on vision-centric questions, questions relying on visual cues only, indicating that these models do not extract useful signal from video frames. Rather than helping, visual input actively disrupts their predictions, pulling them away from the textual priors they would otherwise rely on.

Observation #7: Proprietary models lose their advantage when given only frames. Under full-modality input, proprietary models lead open-source across all question types, but this advantage collapses under frames-only, Table 5. On vision-centric questions, open-source models score better than proprietary (3.15 vs. 3.06). We also observe that open-source models show smaller performance drops under frames-only in comparison to proprietary models. Combined with Observation #6, this suggests that proprietary superiority is driven by stronger language priors, not better visual understanding.

6.4. Performance Across Question Categories

Observation #8: The primary performance bottleneck is visual perception, not reasoning type. Performance across question categories follows a clear pattern: categories that reward dialogue comprehension (CRD, NPA) are easiest for models (3.53, 3.41 proprietary Factuality), while STA (i.e., requiring fine-grained perception of lighting, camera work, and environmental detail) is hardest by a wide margin (2.98 proprietary, 2.83 open-source). Interestingly, humans show no such drop (4.15), making the human–model gap on STA (1.17–1.32) roughly double that of CRD or NPA. The modality ablations sharpen the point: on STA, proprietary models given dialogue-only *improve* by +0.63 Factuality, while those given frames-only *decline* by −0.19, indicating

Model	Question Types			Question Categories				
	Dialogue	Scene	Multimodal	CRD	NPA	STA	TEMP	TH
Relevance Score								
Open-Source Models	3.61	3.53	3.52	3.53	3.52	3.58	3.70	3.72
(only frames)	+0.06	+0.14	+0.15	+0.15	+0.07	+0.08	+0.00	+0.11
(only dialogue)	+0.05	−0.09	+0.08	+0.09	+0.03	−0.14	−0.16	+0.01
Proprietary Models	3.84	3.63	3.83	3.82	3.74	3.53	3.73	3.86
(only frames)	−0.18	+0.01	−0.11	−0.14	−0.09	+0.11	−0.11	−0.17
(only dialogue)	+0.37	+0.24	+0.30	+0.33	+0.32	+0.31	+0.15	+0.35
Human*	4.27	3.97	4.00	4.05	3.98	4.41	–	4.11
Factuality Score								
Open-Source Models	3.21	3.05	3.11	3.19	3.13	2.83	3.15	3.02
(only frames)	−0.14	+0.08	−0.02	−0.07	−0.03	+0.12	−0.06	−0.09
(only dialogue)	−0.01	+0.02	+0.10	+0.07	+0.08	−0.05	−0.16	+0.03
Proprietary Models	3.63	3.15	3.46	3.53	3.41	2.98	3.23	3.38
(only frames)	−0.48	−0.09	−0.25	−0.29	−0.22	−0.19	−0.26	−0.38
(only dialogue)	+0.49	+0.73	+0.70	+0.64	+0.69	+0.63	+0.61	+0.59
Human*	4.17	3.84	3.98	4.07	3.86	4.15	–	4.14

Table 5. **Ablation Experiments.** We report the average performance of an open-source model and a proprietary model on our proposed reference-free metric, broken down by question types and categories. We include ablations where models are prompted with only video frames, only dialogue, or the full context (blue rows). We also report the average performance of *Humans** on each question type and category using the sampled set of 118 questions.

that visual encoders actively hurt on questions demanding the most from them.

7. Conclusion

We proposed MovieRecapsQA, a multimodal open-ended VideoQA benchmark built from aligned recap videos, full-length movies, subtitles, and summaries. By leveraging recap videos as an intermediate representation, our dataset supports questions that require reasoning over both visual and linguistic context while preserving the narrative structure of the underlying film. In addition, we proposed on atomic fact-based, reference-free evaluation framework that scores answers on factuality, coherence, and relevance without relying on a single reference response.

Taken together, our analyses highlight three broader implications for long-form VideoQA:

1. **Metrics matter.** Conventional semantic metrics and reference-based LLM judges are not reliable indicators of factual correctness in open-ended VideoQA. Our fact-based metric better separates models and humans and exposes genuine weaknesses, suggesting that future benchmarks should incorporate similar fact-grounded evaluation rather than relying solely on reference answers.
2. **Multimodality is still brittle.** The visual information gap—where removing frames improves factuality—indicates that current MLLMs do not robustly integrate video and dialogue. Progress on visual encoders alone is not enough; we need architectures and training strategies that force models to ground language in visual evidence rather than treating subtitles as the primary source of truth.
3. **Recap videos are a powerful but underused signal.** By aligning recap summaries, recap videos, and full movies, MovieRecapsQA provides a scalable source of atomic facts and multimodal context that can drive both better models and better evaluation (through fact-based metrics).

We hope these findings encourage future work on (i) architectures that explicitly ground answers in both videos and dialogue, (ii) training objectives that reward visual-textual consistency rather than mere fluency, and (iii) new benchmarks that, like MovieRecapsQA, pair long-form narrative videos with auxiliary textual signals rich enough to support atomic fact extraction and reference-free evaluation.

References

[1] Anthropic. Model card addendum: Claude 3.5 sonnet. https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf, 2024. 6

[2] Kirolos Ataallah, Eslam Mohamed Bakr, Mahmoud Ahmed, Chenhui Gou, Khushbu Pahwa, Jian Ding, and Mohamed Elhoseiny. InfiniBench: A benchmark for large multi-modal models in long-form movies and TV shows. In *Proceedings of the 2025 Conference on Empirical Methods in Natural*

Language Processing, pages 19496–19523, Suzhou, China, 2025. Association for Computational Linguistics. 3

[3] 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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. arXiv:2502.13923 [cs]. 6

[4] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. 3

[5] Meng Cao, Pengfei Hu, Yingyao Wang, Jihao Gu, Haoran Tang, Haoze Zhao, Jiahua Dong, Wangbo Yu, Ge Zhang, Ian Reid, and Xiaodan Liang. Video simpleqa: Towards factuality evaluation in large video language models. *CoRR*, abs/2503.18923, 2025. 2

[6] Seongho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Minsu Lee, and Byoung-Tak Zhang. DramaQA: Character-Centered Video Story Understanding with Hierarchical QA, 2020. arXiv:2005.03356 [cs]. 3, 4

[7] Google DeepMind. Gemini 2.5 flash model card. 2025. 6

[8] Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States, 2022. Association for Computational Linguistics. 3

[9] Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9: 391–409, 2021. 2, 3

[10] Haoqi Fan, Yanghao Li, Bo Xiong, Wan-Yen Lo, and Christoph Feichtenhofer. Pyslowfast. https://github.com/facebookresearch/slowfast, 2020. 3

[11] Tanya Goyal and Greg Durrett. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online, 2020. Association for Computational Linguistics. 3

[12] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. 2025. 2

[13] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the Potential of Small Language Models with Scalable Training Strategies, 2024. arXiv:2404.06395 [cs]. 6

[14] A. et al. Hurst. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6

[15] Amazon Artificial General Intelligence. The amazon nova family of models: Technical report and model card. *Amazon Technical Reports*, 2024. 6

[16] Md Mohaiminul Islam, Mahmudul Hasan, Kishan Shamsundar Athrey, Tony Braskich, and Gedas Bertasius. Efficient movie scene detection using state-space transformers. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18749–18758, 2023. 3, 1, 2

[17] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering, 2017. arXiv:1704.04497 [cs]. 4

[18] Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada, 2023. Association for Computational Linguistics. 2

[19] Ryo Kamoi, Tanya Goyal, and Greg Durrett. Shortcomings of question answering based factuality frameworks for error localization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 132–146, Dubrovnik, Croatia, 2023. Association for Computational Linguistics. 3

[20] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online, 2020. Association for Computational Linguistics. 3

[21] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium, 2018. Association for Computational Linguistics. 3, 4

[22] Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, Chunyuan, and Yuanhan Zhang. Llava-next: A strong zero-shot video understanding model, 2024. 6

[23] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. pages 22195–22206, 2024. 3

[24] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. 3, 5

[25] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, 2023. Association for Computational Linguistics. 3, 5

[26] Wentao Ma, Weiming Ren, Yiming Jia, Zhuofeng Li, Ping Nie, Ge Zhang, and Wenhu Chen. Videoeval-pro: Robust and realistic long video understanding evaluation, 2025. 2

[27] Karttikeya Mangalam, Raiymbek Akshkulakov, and Jitendra Malik. Egoschema: a diagnostic benchmark for very long-form video language understanding. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2023. Curran Associates Inc. 2, 3

[28] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wentau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 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, 2023. Association for Computational Linguistics. 3

[29] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models, 2023. 3

[30] Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online, 2021. Association for Computational Linguistics. 2

[31] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. 3

[32] Sanjana Ramprasad and Byron C Wallace. Do automatic factuality metrics measure factuality? a critical evaluation, 2025. 3

[33] Ruchit Rawal, Khalid Saifullah, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. In *CVPR 2024 Workshop SyntaGen: Harnessing Generative Models for Synthetic Visual Datasets*, 2024. 2, 3, 4

[34] Yixiao Song, Yekyung Kim, and Mohit Iyyer. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474, Miami, Florida, USA, 2024. Association for Computational Linguistics. 3

[35] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. MovieQA: Understanding Stories in Movies through Question-Answering, 2016. arXiv:1512.02902 [cs]. 4

[36] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: a benchmark for long-context interleaved video-language understanding. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 2

[37] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Learning to answer visual questions from web videos, 2022. 3

[38] Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. Helmet: How to evaluate long-context language models effectively and thoroughly. In *International Conference on Learning Representations (ICLR)*, 2025. 3, 5

[39] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, pages 27263–27277. Curran Associates, Inc., 2021. 5

[40] Hongjie Zhang, Yi Liu, Lu Dong, Yifei Huang, Zhen-Hua Ling, Yali Wang, Limin Wang, and Yu Qiao. MoVQA: A Benchmark of Versatile Question-Answering for Long-Form Movie Understanding, 2023. arXiv:2312.04817 [cs]. 3, 4

[41] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. 3, 5