DRAMAVQA: PROBING DETAILED UNDERSTANDING OF ULTRA-LONG TV DRAMAS

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite recent progress in long-form video understanding, current benchmarks mostly focus on short clips or hour-scale videos. This leaves a major gap in evaluating ultra-long video understanding, where videos often last tens of hours. To bridge this gap, we introduce DramaVQA, a large-scale benchmark built upon long-form TV dramas. Containing approximately 20K question-answering (QA) pairs, the scale of DramaVQA exceeds previous long-video datasets, such as Video-MME and LVBench, by an order of magnitude. Our benchmark has two main features that make it stand out from existing datasets. First, we establish a scalable, LLM-assisted annotation pipeline that uses multimodal large language models (MLLMs) to generate complex questions, followed by strict human checking to ensure high data quality. Second, DramaVQA introduces a multi-level time structure; queries are designed across different video lengths, with the hardest questions requiring overall understanding across the entire drama. Extensive evaluations show that state-of-the-art MLLMs face major challenges when tested on these very long videos, highlighting the need for agentic AI algorithms capable of managing long-term memory. *We will release all data and code to facilitate further research.*

1 INTRODUCTION

The rapid evolution of multimodal large language models (MLLMs) has greatly expanded the limits of video understanding. Recent state-of-the-art models, such as Gemini-3-Pro Team et al. (2023) and the Qwen-VL series Bai et al. (2025b;a), now support very large context windows (*e.g.*, 128K input tokens). This allows them to achieve remarkable accuracy on question-answering (QA) tasks for hour-long videos previously considered impossible. However, as model capacities grow, the community lacks a benchmark that pushes the frontier toward ultra-long contexts, specifically videos lasting tens of hours, where processing the entire video at once is no longer computationally practical.

To address this gap, we introduce DramaVQA, a large-scale benchmark built upon long-form TV dramas to encourage research in ultra-long video understanding. TV dramas serve as an ideal testbed because they feature continuous, detailed storylines driven by character growth and long-range cause-and-effect relationships. This structure allows us to create challenging questions that require finding key information across dozens of episodes. As a result, it rigorously tests MLLM capabilities across multiple dimensions, including visual recognition, temporal grounding, and high-level summarization.

The main hurdle in creating such a benchmark is the high cost of human annotation and the difficulty of ensuring quality control. To overcome this, we propose a scalable, LLM-assisted annotation pipeline with four key stages: (1) extracting structured metadata from the dramas (*e.g.*, subtitles, facial identities, speaker identities) and using MLLMs to generate multi-level video captions; (2) using LLMs to generate high-quality candidate QA pairs from these captions; (3) employing human annotators to carefully filter and correct wrong or ambiguous QAs; and (4) using LLMs one last time to verify that the remaining questions are unique across the entire drama. This method allows DramaVQA to reach an unprecedented scale of approximately 20K QA pairs, an order of magnitude larger than existing long-video datasets like Video-MME Fu et al. (2025) and LVBench Wang et al. (2025a). Table 1 shows a detailed comparison between DramaVQA and recent benchmarks.

This construction process naturally gives DramaVQA a multi-level time structure. Each QA pair can be tested using different input video lengths, categorized into four distinct temporal tracks: `Medium` (averaging 7.6 minutes), `Long` (averaging 43.1 minutes), `Extra-Long` (averaging 9.3 hours), and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Benchmark	#QAs [per video]	Avg Dura.	MG	VM	TG	Annotation
TVQA-test Lei et al. (2018)	15,253 [1.0]	0.003h	✗	✗	–	Manual
ActivityNet-QA Yu et al. (2019)	800 [1.0]	0.03h	✗	✗	–	Manual
MVBench Li et al. (2024c)	4,000 [1.1]	0.04h	✗	✗	–	Auto
EgoSchema Mangalam et al. (2023)	5,063 [1.0]	0.05h	✗	✗	–	Auto+Manual
MovieQA Tapaswi et al. (2016)	6,462 [1.0]	0.06h	✗	✗	–	Manual
LongVideoBench Wu et al. (2024)	6,678 [1.8]	0.13h	✗	✓	–	Manual
MovieChat-1K Song et al. (2024)	13,000 [13.0]	0.16h	✗	✗	–	Manual
MoVQA Zhang et al. (2025b)	20,061 [200.6]	0.28h	✓	✗	✗	Manual
CG-Bench Chen et al. (2024a)	12,129 [10.0]	0.45h	✓	✓	✓	Manual
Video-MME-L Fu et al. (2025)	900 [3.0]	0.69h	✗	✓	✗	Manual
LVBench Wang et al. (2025a)	1,549 [15.0]	1.14h	✗	✗	✗	Manual
DramaVQA-UL	**20,000** [**1,666.7**]	**38.85h**	✓	✓	✓	Auto+Manual

Table 1: **DramaVQA** excels in **QA number**, **average duration**, and other attributes among recent long-form video benchmarks. **MG**: multi-granular (each QA supports different durations); **VM**: video metadata available (subtitles, faces, speaker data, hierarchical video captions); **TG**: temporal grounding provided (only for 0.2h+ videos).

`Ultra-Long` (the full drama, averaging 38.9 hours). Notably, the latter two tracks are much longer than existing benchmarks (*e.g.*, the average durations of Video-MME-Long and LVBench are only 0.69 and 1.14 hours, respectively). Our extensive evaluation of state-of-the-art MLLMs shows a significant performance drop as the video length increases. While current models perform reasonably well on the `Medium` track, their accuracy drops significantly on the `Long` and track, and falls to near-random guessing levels on the `Extra-Long` and `Ultra-Long` tracks. This sharp drop shows a major weakness in current architectures: they rely too heavily on brute-force context scaling and cannot handle ultra-long videos.

Beyond providing a strict testbed for zero-shot video QA, DramaVQA pushes the field forward in another important way. First, we will release over 400K high-quality QA pairs. Although many were filtered from the final benchmark for failing the strict uniqueness test, all these QAs have been human-validated and include precise timestamps, offering a highly valuable, large-scale dataset for MLLMs. Unlike traditional benchmarks that provide only 3–20 questions per video, DramaVQA supplies an average of 1,600+ questions per drama. This high density of queries over massive time scales challenges the community to move beyond simple, single-pass processing. Instead, it motivates a necessary shift toward memory-efficient long video understanding, where models must organize video streams into compact, easily searchable memory formats to perform complex reasoning within practical computational limits.

2 Related Work

Multimodal large language models (MLLMs). The field of computer vision has greatly changed with the rise of MLLMs Hurst et al. (2024); Chen et al. (2024b); Wang et al. (2024a); Bai et al. (2025b;a). By building on the strong abilities of basic LLMs Touvron et al. (2023); Achiam et al. (2023); Team et al. (2023); Guo et al. (2025); Yang et al. (2025), these models convert visual signals into discrete tokens, easily combining them with text prompts Liu et al. (2023); Zhu et al. (2023); Kim et al. (2024); Tong et al. (2024). This step has moved the field past simple, closed-set tasks (*e.g.*, basic detection or image classification) into open-ended video question answering Maaz et al. (2024); Lin et al. (2024); Zhang et al. (2024b); Li et al. (2024b;a). Furthermore, ongoing model improvements allow them to process much larger sequences of visual tokens, opening up new ways to understand longer videos.

Benchmarks for long-form video understanding. Since current models have largely mastered static images and short video clips, research has shifted toward longer video analysis. Recent large-scale datasets focused on hour-long videos and complex QA tasks (*e.g.*, EgoSchema Mangalam et al. (2023), LongVideoBench Wu et al. (2024), Video-MME Fu et al. (2025), LVBench Wang et al. (2025a), CG-Bench Chen et al. (2024a)) have created major challenges for modern vision systems. The main trend among these datasets is the steady increase of the input length (as shown in Table 1), which challenges MLLMs to integrate massive amounts of video information into their context windows. Meanwhile, some benchmarks Cheng et al. (2025); Liu et al. (2025) use minute-level videos paired with complex

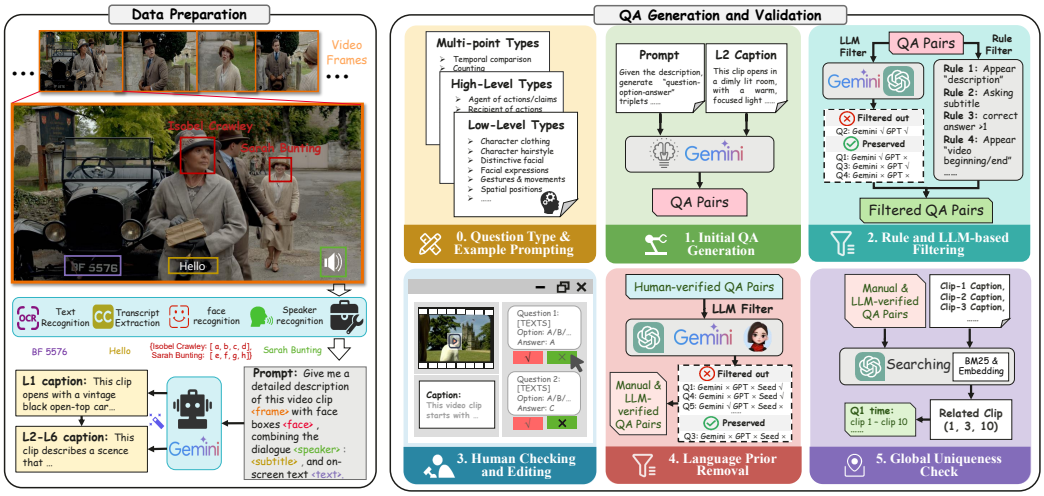

Figure 1: An illustration of the two-stage procedure to establish the DramaVQA benchmark.

logical questions to test the reasoning limits of these models. Although LongTimeScope Zohar et al. (2025) and Ego-R1 Bench Tian et al. (2025) cover much longer videos, they contain only 450 and 300 evaluation questions, respectively, with limited coverage of fine-grained understanding within individual videos. To address this, our research delivers the largest benchmark built to date for rigorous, ultra-long video reasoning, constructed specifically from long-form TV dramas. While datasets like MovieQA Tapaswi et al. (2016), MovieChat Song et al. (2024), DramaQA Choi et al. (2021), and MoVQA Zhang et al. (2025b) also focus on video understanding using movie data, their video durations are still at the sub-hour level. By using multi-episode TV dramas, we make it possible to extend the video duration to the 10-hour level.

Advanced solutions for long-form video understanding. To mitigate inherent context limitations of MLLMs, existing methods fall into two categories: video compression that condenses visual content for efficient context utilization Li et al. (2024d); Lee et al. (2024); Shen et al. (2024a;b); Choudhury et al. (2024); Zhang et al. (2024a), and chunk-based processing that segments long videos for step-wise inference Song et al. (2026); Tang et al. (2025b); Zhang et al. (2025a). Integrated with powerful reasoning models, chunking solutions have evolved into agentic frameworks, dominating mainstream benchmarks with rational reasoning and tool scheduling Hurst et al. (2024); Guo et al. (2025); Yang et al. (2025); Fan et al. (2024); Wang et al. (2024b; 2025c); Feng et al. (2025); Tang et al. (2025a); Qiu et al. (2026). Targeting extreme ultra-long video scenarios, DramaVQA further motivates robust agentic designs and memory-based mechanisms to address long-range video understanding challenges Wang et al. (2025b); Diko et al. (2025); Jin et al. (2025); Zuo et al. (2025).

3 The DramaVQA Benchmark

The overall pipeline for building the DramaVQA benchmark is shown in Figure 1. This process consists of two main stages: (1) data preparation, where we extract structured metadata from the TV dramas and use an MLLM to generate detailed video descriptions across different time levels; and (2) QA generation and validation, which combines LLM generation with strict human checking to ensure both the large scale and the high quality of the dataset.

3.1 Data Preparation

Data collection. Our dataset comprises 15 long-form TV series spanning multiple seasons, with a total runtime of 606 hours. Specifically, DramaVQA includes 12 totalling 466 hours, and the other three series are reserved for the training split. We extract metadata through a four-stage pipeline. We process each drama $\mathcal{V}$ individually; for simplicity, denote T as its total duration, and omit the drama index for the rest of this work. Details of this part can be found in Appendix A.

1. **Transcript and text extraction.** We use PaddleOCR Cui et al. (2025b;a) followed by Qwen3-VL-32B Bai et al. (2025a) to extract and refine hardcoded subtitles and in-screen texts. This produces a sequence of utterances $\mathcal{U} = \{u_1, \ldots, u_N\}$, where each $u_n = (\mathbf{w}_n, [s_n, e_n])$ contains the text $\mathbf{w}_n$ and start/end timestamps $s_n, e_n \in [0, T]$, and a set of texts $\mathcal{T} = \{t_1, \ldots, t_M\}$, where each $t_m = (\mathbf{w}_m, [s_m, e_m])$.
2. **Character library construction.** We build a cast roster $\mathcal{C}$ from credits and web data. We initialize a facial registry with official portraits and enrich it using 1-FPS video sampling to capture varied appearances. Then, a frame-by-frame processing is performed to detect all facial instances $\mathcal{F} = \{f_1, \ldots, f_R\}$. Here, each $f_r = (c_r, t_r, \mathbf{b}_r)$ links a detected face to an identity $c_r \in \mathcal{C}$ at timestamp t_r and bounding box $\mathbf{b}_r$.
3. **Speaker label annotation.** We combine audio features within $[s_n, e_n]$ and visual instances $\mathcal{F}$ to assign initial speaker labels using a label propagation algorithm to find 'who spoke when'. Human labelers then perform a sentence-by-sentence check following a well-established guideline, updating each utterance u_n with a final speaker identity $c_n \in \mathcal{C}$.
4. **Hierarchical video captioning.** We use a multi-level temporal structure (L1–L6) to summarize the long video. First, the video is split into 10–15s base units (called L1 units) $h_k^{(1)} \in \mathcal{H}^{(1)}$ using the PySceneDetect algorithm Castellano (2024) with CLIP similarity Radford et al. (2021). For these units, Qwen3-VL-32B Bai et al. (2025a) generates visual descriptions $\mathbf{d}_k^{(1)}$. Next, we cluster these into L2 units and use Qwen3-32B Yang et al. (2025) to summarize them. This repeats across six levels, $\mathcal{H}^{(l)}$ for $l \in \{1, \ldots, 6\}$. Each segment is formalized as $h_k^{(l)} = (\mathbf{d}_k^{(l)}, [s_k^{(l)}, e_k^{(l)}])$, with start and end timestamps $s_k^{(l)}$ and $e_k^{(l)}$. This ends with a single L6 summary for the full drama $\mathcal{V}$.

3.2 QA Generation and Validation

We design a sophisticated pipeline to generate QA pairs from the multi-level captions, based on three main rules: (i) **diversity**, ensuring a wide coverage of question types; (ii) **reducing language prior**, making sure models cannot answer questions using only their internal knowledge (like drama scripts in their training data) without seeing the video; and (iii) **global uniqueness**, ensuring questions are not repeated across the entire drama to test ultra-long reasoning. To prevent the data from favoring any specific model, we use several different LLMs across the different stages. Details of all these steps are in Appendix B.

1. **Initial QA generation.** We use Gemini-2.5-Pro to generate initial QA pairs from L2 captions $h_k^{(2)}$. These segments, with an average length of 1.2 minutes, offer a good balance between visual details and the main storyline. Following previous methods (*e.g.*, LLaVA-Video-178K Zhang et al. (2024b)), we ask for different question types (with demonstrations) in the prompt, including visual perception, spatial understanding, action & interaction, event understanding, temporal understanding, relation & reasoning, and counting. All entries are single-choice questions with 2–8 (average 6+) options.
2. **Rule and LLM-based filtering.** To reduce hallucination and internal knowledge prior, we first use an LLM (Qwen3-32B) to remove 'self-answering' questions where the answer is already in the question text, as well as invalid questions that rely on segment-specific time clues (*e.g.*, 'at the beginning of the clip'). In addition, two LLMs (GPT-5 and Gemini-2.5-Pro) are asked to answer the questions based only on the text options; any QA answered correctly by both models is removed.
3. **Human checking and editing.** We recruit human labelers to check the remaining questions. Labelers are assigned L2 segments and asked to watch the video $h_k^{(2)}$ to make sure each QA meets three rules: (a) the question matches the video and is concise; (b) the choices are clear; (c) the answer is correct and unambiguous; and (d) the question is not repeated. We implement a two-stage labeling procedure, where the second pass rigorously double-checks and improves the quality of the first pass.
4. **Language prior removal.** After human editing, we give the QAs an even stricter text-only test using three models: GPT-5, Gemini-2.5-Pro, and Seed1.8 Pro. To further increase the difficulty, we only keep QA pairs that are answered incorrectly by all three LLMs. This high standard ensures a strong gap between text-only and vision-language results.

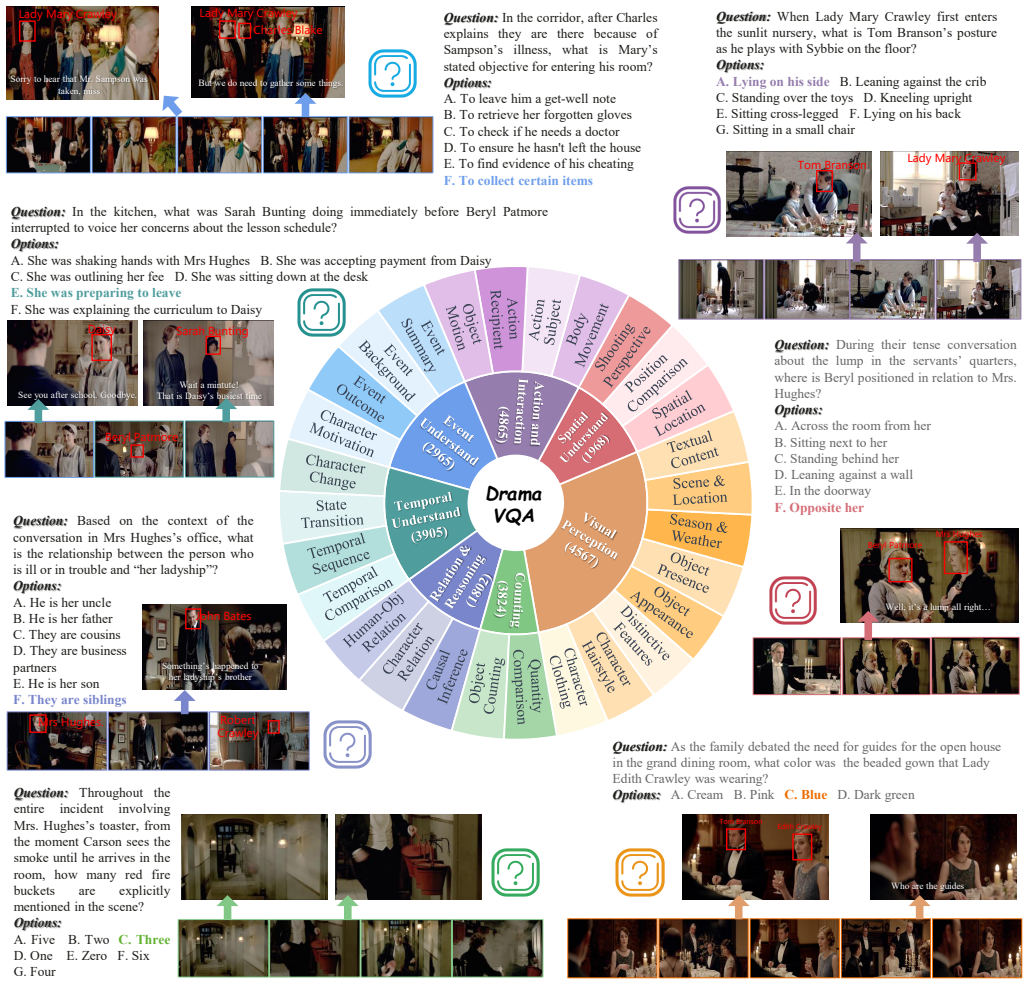

Figure 2: Two-level QA categorization and representative examples (one for each Level-1 category; the color of each question logo corresponds to that of the category sector). We also display the keyframes that the model must sample for question-answering.

5. **Global uniqueness check.** We use a global search algorithm to make sure each QA is unique across the entire drama length T. We combine language embeddings from the bge-large-en/zh-v1.5 model Xiao et al. (2024) and the BM25 algorithm Robertson and Zaragoza (2009) to find similar L2 segments. Fo the top-ranked similar segments, an LLM extracts keywords from the QA; if another L2 segment contains these keywords, the QA is marked as not unique and removed. This ensures that the question can only be answered by referring to the specific target video segment, thereby improving the uniqueness and discriminative power of the QA pairs.

The dataset contains 25,404 L2 segments, of which 20,278 belong to DramaVQA. After the first four steps, the average numbers of QA pairs per L2 segment are 103.1, 27.3, 18.8, and 5.4, respectively. The training set retains all QA pairs (about 15k) generated after Step 4. These pairs are then processed in Step 5 to annotate the uniqueness level of each question. The benchmark set retains only QA pairs that are globally unique within their entire drama after Step 5, resulting in 1.7 QAs per L2 segment. We then sample the final 20,000 questions to ensure balanced distributions across question types and video timestamps.

3.3 Benchmark Statistics

We have established 20K QA pairs across 12 TV dramas, representing a scale one order of magnitude larger than existing long-video benchmarks. Furthermore, the filtered QAs (400K+), while not unique

throughout the entire drama, can also be used as training data and served as a high-quality resource for supervised fine-tuning and advancing long-form video understanding. An illustration of seven basic categories and their breakdown into 28 sub-categories is shown in Figure 2. Detailed dataset statistics are provided in Appendix A (Table 4).

Because each preserved question is verified to be unique throughout the entire drama duration T, DramaVQA supports **multi-granular evaluation**. In our experiments, we use L3 through L6 segments to define four distinct evaluation tracks: `Medium` (averaging 7.6 minutes), `Long` (averaging 43.1 minutes), `Extra-Long` (averaging 9.3 hours), and `Ultra-Long` (the full drama, averaging 38.9 hours). Notably, our `Medium` and `Long` tracks are similar in length to the 'long-form' subsets of existing benchmarks, while the `Extra-Long` and `Ultra-Long` tracks present an unprecedented challenge for modern MLLMs.

Another interesting feature of our benchmark is the **high degree of inter-connectivity** between QA pairs within a single drama. For example, the answer to one question may serve as a condition for another. These connections enable us to construct complex tasks that require multi-hop reasoning or retrieving and integrating multiple pieces of information, providing a rigorous assessment of MLLMs' reasoning and multi-point information retrieval capabilities. We present a preliminary study of these complex reasoning tasks in Section 4.4.

4 Experiments

4.1 Models and Evaluation Metrics

Models. We evaluate two main groups of models to build a strong baseline. The first group includes state-of-the-art MLLMs, such as Qwen3-VL-8B/32B Bai et al. (2025a), Qwen3.5-27B Team (2026a), Qwen3.6-27B Qwen Team (2026), GLM-4.6V Hong et al. (2025) and GLM-5.3-Flash-FP8 Team (2026b). These models uniformly sample a set number of frames (256 or 512) from the input video, tokenize them as visual context, and then perform inference. We strictly follow the default settings for these models; Thinking settings are specified in Table 2. The second group features recently released, agentic MLLMs, namely Conan Ouyang et al. (2025) and LongVideo-R1 Qiu et al. (2026), which start with a basic video input and dynamically sample extra frames for question-answering.

Furthermore, we test a simple baseline named Naive-RAG that searches through the clip captions generated across the video. This method shares the same input structure as LongVideo-R1 Qiu et al. (2026): for each input video, we build a 3-level tree structure, where each clip is evenly divided into $D = 6$ smaller clips at the next level. We extract a caption for every clip and use BM25 search to find the top-3 most relevant captions at each level. Finally, these 9 captions are combined and fed into an LLM to answer the question. We also develop a beam-search variant (RAG-BS) where, at the second and third levels, the search is strictly limited to the child nodes of the top-3 captions chosen in the previous level.

Data. We test the models on the entire set of 20K QA pairs created in the previous section. As explained earlier, each question is tested using different lengths of video input. These correspond to four distinct tracks: `Medium`, `Long`, `Extra-Long`, and `Ultra-Long`. To make it easier for the community to reproduce our results, we also build a 'mini' subset containing about 5% of the main test set. On this smaller set, we test tree proprietary models, Seed2.0-Pro Seed (2026), GPT-5.5 OpenAI (2026) and Gemini-3.5-Flash Team et al. (2023), and compare them against the open-source models.

Evaluation Protocol.

For each sampled frame, original hardcoded subtitles are retained, forcing models to read text directly from images. To aid character recognition, we draw red bounding boxes for detected faces and add name labels. Models are required to output a single option (e.g., `[A]`). For each QA pair, we repeatedly prompt MLLMs for up to five attempts. If no valid answer is produced, we assign a random-guess score of $1/X$, where X denotes the number of options. We report both QA accuracy and computational overhead[1]. Notably, the overhead of agentic methods, including RAG, covers video captioning costs. Full experimental settings are provided in Appendix C.1.

[1]All models run on an 8×A800 GPU server with vLLM. Latency is measured in single-thread mode, while multi-thread testing brings 6–10× speedup.

Model	Fr.	Th.	Medium (L3) Mini [Full] (time)	Long (L4) Mini [Full] (time)	Extra-long (L5) Mini [Full] (time)	Ultra-long (L6) Mini [Full] (time)
Human	–	–	81.0 *(215)*	–	–	–
No-vision Models						
Random Guess	0.0	✗	16.2 [16.2] (0.0)	16.2 [16.2] (0.0)	16.2 [16.2] (0.0)	16.2 [16.2] (0.0)
Language Only (Q3)	256	✗	21.5 [22.2] (0.1)	21.5 [22.2] (0.1)	21.5 [22.2] (0.1)	21.5 [22.2] (0.1)
Language Only (Q3.5)	256	✓	23.1 [23.3] (33.9)	23.1 [23.3] (33.9)	23.1 [23.3] (33.9)	23.1 [23.3] (33.9)
Proprietary Models						
Seed2.0 Pro Seed (2026)	256	✓	73.2 (48.5)	49.2 (63.4)	32.2 (60.1)	30.7 (49.2)
GPT-5.5 OpenAI (2026)	256	✗	67.2 (64.6)	45.2 (55.5)	29.9 (81.0)	27.6 (56.1)
Gemini-3.5-Flash Team et al. (2023)	256	✗	77.5 (38.5)	61.2 (36.4)	43.9 (36.8)	37.1 (56.0)
Open-source Native MLLMs						
GLM-4.6V-106B Hong et al. (2025)	256	✗	51.0 (15.4)	34.5 (14.8)	27.4 (14.3)	26.5 (12.6)
GLM-4.6V-106B Hong et al. (2025)	256	✓	57.1 (19.5)	36.6 (18.5)	27.1 (16.1)	25.4 (15.1)
GLM-5.3-Flash-FP8 Team (2026b)	256	✓	70.5 (31.7)	49.9 (14.4)	34.9 (14.3)	32.1 (14.2)
Qwen3-VL-32B Bai et al. (2025a)	256	✓	62.0 (20.5)	37.9 (18.3)	26.4 (14.9)	24.9 (16.8)
Qwen3.5-27B Team (2026a)	256	✗	69.4 (26.3)	47.1 (16.6)	35.7 (13.2)	34.4 (12.2)
Qwen3-VL-8B Bai et al. (2025a)	256	✗	40.9 [42.3] (17.9)	28.7 [30.2] (11.5)	24.8 [25.1] (13.0)	22.7 [24.1] (11.2)
Qwen3-VL-8B Bai et al. (2025a)	256	✓	54.2 [56.3] (19.2)	36.5 [37.7] (17.6)	24.6 [26.7] (16.3)	23.4 [24.9] (12.6)
Qwen3-VL-32B Bai et al. (2025a)	256	✗	51.0 [51.3] (22.6)	35.1 [35.0] (17.9)	27.8 [27.4] (18.2)	26.4 [25.9] (17.4)
Qwen3.6-27B Qwen Team (2026)	256	✗	69.9 [67.1] (50.5)	48.0 [49.5] (23.8)	35.2 [36.9] (22.5)	33.0 [34.6] (20.1)
Open-source Native MLLMs						
GLM-4.6V-106B Hong et al. (2025)	512	✗	55.0 (28.0)	37.0 (29.3)	27.4 (31.1)	26.6 (30.6)
GLM-4.6V-106B Hong et al. (2025)	512	✓	60.8 (34.7)	40.1 (34.6)	25.6 (35.0)	24.6 (37.5)
GLM-5.3-Flash-FP8 Team (2026b)	512	✓	75.4 (28.2)	56.1 (33.5)	36.1 (31.3)	32.1 (32.6)
Qwen3-VL-32B Bai et al. (2025a)	512	✓	66.1 (28.6)	43.7 (33.3)	26.4 (30.7)	24.1 (21.3)
Qwen3.5-27B Team (2026a)	512	✗	72.2 (27.4)	51.0 (27.6)	37.1 (26.4)	33.9 (23.8)
Qwen3-VL-8B Bai et al. (2025a)	512	✗	41.7 [42.5] (23.3)	30.4 [31.1] (26.0)	24.1 [25.4] (26.8)	22.6 [24.2] (29.0)
Qwen3-VL-8B Bai et al. (2025a)	512	✓	57.2 [58.9] (29.8)	40.0 [41.0] (31.5)	26.9 [27.5] (32.2)	25.0 [25.2] (26.1)
Qwen3-VL-32B Bai et al. (2025a)	512	✗	54.8 [53.3] (33.3)	35.8 [37.2] (36.4)	28.2 [27.5] (37.6)	25.4 [26.2] (37.9)
Qwen3.6-27B Qwen Team (2026)	512	✗	72.5 [70.8] (41.4)	55.0 [54.9] (40.9)	37.9 [38.6] (36.8)	34.1 [35.1] (30.4)
Open-source Agentic MLLMs						
Conan Ouyang et al. (2025)	256	✓	48.8 [50.1] (63.3)	36.3 [37.1] (60.5)	29.7 [30.9] (59.7)	28.6 [29.9] (68.3)
Conan Ouyang et al. (2025)	512	✓	49.0 [48.8] (87.5)	37.7 [37.9] (79.5)	29.7 [30.5] (79.6)	29.6 [29.5] (84.1)
RAG-BS (Q3)	var	✗	47.4 [47.1] (180.4)	33.5 [33.7] (372.1)	25.8 [25.7] (408.2)	24.8 [24.7] (406.5)
Naive-RAG (Q3)	var	✗	51.7 [50.4] (292.9)	46.6 [45.9] (1664.5)	29.7 [29.1] (1955.5)	26.3 [25.3] (1955.0)
RAG-BS (Q3.8)	var	✗	56.4 [56.0] (180.5)	40.4 [41.1] (372.2)	32.1 [32.7] (408.3)	30.9 [31.5] (406.5)
Naive-RAG (Q3.8)	var	✗	60.1 [59.8] (293.0)	56.7 [56.1] (1664.6)	36.9 [36.8] (1955.5)	31.5 [32.7] (1955.0)
LongVideo-R1 Qiu et al. (2026)	var	✓	62.8 [63.7] (70.1)	54.8 [55.3] (90.1)	34.7 [34.0] (98.3)	27.2 [26.6] (93.2)
LongVideo-R1 (post-tuned on our data)	var	✓	66.1 [67.5] (81.7)	60.0 [61.4] (113.7)	40.0 [40.9] (163.8)	30.8 [30.5] (172.0)

Table 2: QA accuracy (%) and average Mini inference time on DramaVQA. Each cell reports Mini accuracy, Full accuracy in square brackets when available, and Mini time (seconds) in parentheses. **Fr.**: sampled frames; **Th.**: thinking enabled; **var**: varying frame counts. Times include retries (at most five attempts per question); more frequent failures can make smaller models slower. Q3, Q3.5, and Q3.8 denote Qwen3-VL-32B, Qwen3.5-27B, and Qwen3.8-27B, respectively. Proprietary models are evaluated only on Mini; their timing may include network latency.

4.2 Main Results

Table 2 summarizes all test results. We first discuss the available results on the **full-set** that contains all 20,000 QA pairs. The key observations are listed below.

The foundational QA tasks remain challenging. Even within the `Medium` track, the tested models leave substantial room for improvement. Specifically, Qwen3.6-27B Qwen Team (2026) achieves the highest reported full-set accuracy at 70.8% with 512 frames, but its accuracy falls to 35.1% on the `Ultra-Long` track. This difficulty comes from three main factors: (i) our data preparation pipeline adds dense information into the video captions, allowing us to create difficult questions that require deep reasoning and an understanding of complex character relationships; (ii) most questions have 6–8 highly competitive (yet clear) options; and (iii) we removed simpler questions (those solvable via language priors) through our strict text-only filtering step.

Performance decreases as the input video duration increases. The average accuracy on the `Long` track is noticeably lower than that on the `Medium` track, and performance on the `Extra-Long` and `Ultra-Long` tracks drops close to the random guess baseline. This trend matches our expectation: as the video length grows, the chance of a model successfully sampling the key frames containing the answer becomes much lower. This also explains why all models show improved accuracy when the context window increases from 256 to 512 sampled frames. However, even with 512 frames, the performance remains close to the random baseline in the `Ultra-Long` track.

Stronger base MLLMs produce better results. Under the same 512-frame, non-thinking setting, `Medium` accuracy increases from 42.5% for Qwen3-VL-8B to 53.3% for Qwen3-VL-32B Bai et al. (2025a) and 70.8% for Qwen3.6-27B Qwen Team (2026). The corresponding `Ultra-Long` scores

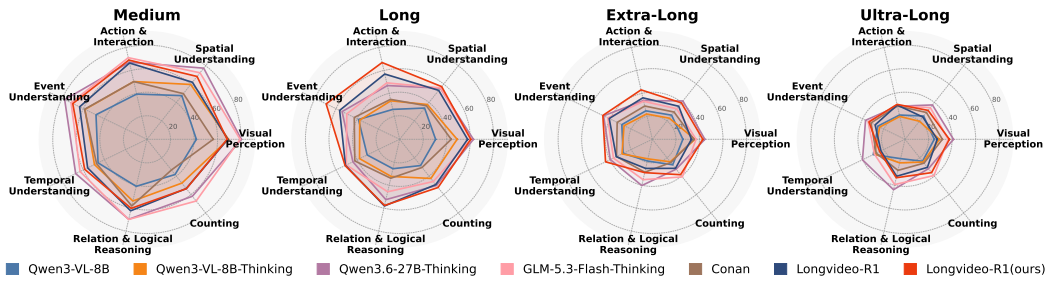

Figure 3: Comparison of QA accuracy (%) with respect to seven major QA categories.

are 24.2%, 26.2%, and 35.1%, respectively. These results support DramaVQA as a useful testbed for evaluating foundation models. However, the smaller absolute gains on the longest track indicate that stronger base models still face substantial limitations when processing extended video contexts.

Thinking mechanisms help, but their benefits depend on video length. For Qwen3-VL-8B with 512 frames, thinking raises `Medium` accuracy from 42.5% to 58.9% and `Long` accuracy from 31.1% to 41.0%. However, the gain on the `Ultra-Long` track is smaller, from 24.2% to 25.2%. Similar trends hold with 256 frames, where thinking yields larger improvements on the shorter tracks. These results suggest that reasoning helps models interpret evidence, while offering limited benefits when sparse sampling fails to capture the information required to answer a question.

Agent-based and retrieval-augmented algorithms show better accuracy but may require higher computational costs. The basic RAG approach, which searches across all short-clip captions, helps reduce the information loss caused by uniform frame sampling. However, this comes at the cost of increased time complexity, which grows with the input video length. Smarter agentic models (*e.g.*, LongVideo-R1 Qiu et al. (2026)) offer a better approach by dynamically sampling a small number of video clips. Still, they cannot always guarantee finding the exact time segments needed for the correct answer. This shows a clear trade-off between accuracy and computational cost, which we discuss further in Section 4.3.

Results on the mini-set. Its performance trends closely align with the full dataset, validating the mini-set as a reliable, cost-effective proxy for fast experiments such as parameter tuning and prompt design. Even proprietary models struggle with ultra-long contexts: while Gemini-3.1-Pro outperforms open-source native MLLMs, its accuracy still declines markedly on `Long` and `Ultra-Long` tracks, mainly due to severe information loss from sparse frame sampling across extended videos.

RL on our training set. To assess the effectiveness of our training data in improving information retrieval and localization across video scales, we adopt the agent-based LongVideo-R1 Qiu et al. (2026) as our baseline. Starting from its original checkpoint, we perform reinforcement learning on our training set. The resulting model outperforms the original LongVideo-R1 by at least 3.6% across all four evaluation tracks, demonstrating the effectiveness of our training data. RL training increases caption retrieval and reasoning activity, with average medium-caption calls rising from 2.38 to 4.65, low-caption calls from 3.12 to 4.66, and reasoning steps from 7.41 to 10.94 per question.

Summary of Table 2. With extensive experiments, we show that ultra-long video understanding remains a major challenge, especially under limited computational budgets. By showing the current limits of existing models, DramaVQA offers a valuable testbed to help guide the future development of state-of-the-art MLLMs and agentic video understanding algorithms.

On QA categories. Figure 3 shows the accuracy distribution across different question categories. Notably, certain question categories (*e.g.*, temporal understanding and counting) are more difficult than others (*e.g.*, visual perception and event understanding). This difference happens mainly because the harder categories require complex logical reasoning or collecting pieces of information scattered across the video. See Appendix C.2 for further analysis.

Failure analysis. Our analysis reveals three main causes for these errors. First, many mistakes happen because sparse frame sampling often misses important keyframes. Second, current models struggle to understand the overall story logic and cause-and-effect relationships. Finally, the lack

Model	Multi-Point (1,510)				Multi-Hop (1,200)	
	TTC-2	TTC-3	CECI	Average	Single-QA	Multi-Hop-QA
Random	50.0	16.7	50.0	39.6	13.7	16.0
Qwen3-VL-8B Bai et al. (2025a)	47.4	18.6	49.2	38.2	39.5	38.4
Qwen3-VL-32B Bai et al. (2025a)	46.4	19.9	43.4	36.1	44.8	42.2
GLM-4.6V-106B Hong et al. (2025)	47.4	18.8	48.2	37.8	48.4	46.0
Qwen3.5-27B Team (2026a)	66.3	57.9	50.6	57.3	61.0	56.8
Qwen3.6-27B Qwen Team (2026)	**81.1**	**70.9**	**59.0**	**69.0**	**67.4**	**63.1**

Table 3: Evaluation on multi-point and multi-hop questions. *Single-QA* denotes the base question taken from the second hop of the paired *Multi-Hop-QA* sample, with identical video input. Parenthesized values are the total number of questions per task.

of audio inputs makes it hard for models to answer dialogue-based questions, such as identifying 'who claimed what'. This clearly shows the need to develop advanced tri-modal (audio-visual-text) MLLMs in the future. More examples are provided in Appendix C.3.

4.3 Accuracy-Efficiency Tradeoff

We compare native MLLMs and agentic algorithms from the perspective of the accuracy-efficiency trade-off. In brief, while agentic methods introduce higher preprocessing cost in case-wise evaluation, they become substantially more competitive in batch-wise settings by sharing captioning cost across many questions from the same long video context. We also observe a practical trade-off between stronger reasoning performance and additional inference overhead. Detailed quantitative results, the full comparison table, and time-cost estimation are provided in Appendix D.

4.4 Multi-Point and Multi-Hop Reasoning: A Preliminary Study

We extend DramaVQA with multi-point and multi-hop questions composed from basic QA pairs to evaluate cross-segment evidence integration and compositional reasoning.

Question construction. Multi-point questions cover *Two-/Three-Event Temporal Comparison* (TTC-2/3) and *Cross-Event Character Identity* (CECI). We sample events from distinct L3 segments within a drama and use templates to query their temporal order or whether characters identified by two base questions are the same. An LLM refines the wording, and we discard questions answered correctly by at least two of three text-only LLMs. For multi-hop questions, we pair questions whose first answer (typically a character name) appears in the second question. An LLM replaces this answer in the second question with a description derived from the first; annotators verify correctness and the necessity of the intermediate answer. See Appendix E for detailed construction procedures, illustrative examples, and supplementary experimental analyses.

Evaluation. We uniformly sample 256 frames from the concatenated L3 segments associated with each composed question. Table 3 shows that Qwen3-VL and GLM-4.6V perform near or below chance on TTC-2 and CECI, and only slightly above chance on TTC-3. Despite stronger temporal comparison by Qwen3.5/3.6, CECI accuracy peaks at 59.0%. All models perform worse on multi-hop questions than on their paired second-hop questions, highlighting reasoning challenges beyond atomic QA. Together, these results reveal limitations in cross-segment evidence integration and compositional reasoning that atomic QA evaluation alone may not fully capture.

5 Conclusions

In this paper, we introduced DramaVQA, a large-scale benchmark designed to advance ultra-long video understanding. Built from long TV drama series, DramaVQA contains about 20K high-quality QA pairs and evaluates models from `Medium` clips to `Ultra-Long` videos lasting tens of hours. Our results show that current MLLMs still struggle on `Extra-Long` and `Ultra-Long` tracks, revealing clear limitations in long-range video reasoning. We also explored the accuracy-efficiency trade-off, finding that agent-based algorithms offer a better balance when answering many questions over long videos. Moving forward, we believe future research must shift from simply expanding context windows toward developing smart memory systems and video-driven world models that can efficiently manage and reason over long-term temporal contexts.

AI USE STATEMENT

We used models from the Gemini, GPT, Doubao, and Qwen families to assist with QA generation and filtering during dataset construction (see Section 3.2 for details), followed by human verification and editing as described in the paper. We also used LLMs to polish the manuscript's language.

ETHICS STATEMENT

DramaVQA focuses on understanding fictional characters, their interactions, and narrative events in television dramas. All questions undergo human review to identify and exclude ethically inappropriate content. The benchmark is intended for research purposes, and we encourage its responsible use.

REPRODUCIBILITY STATEMENT

We provide detailed documentation and supporting artifacts to facilitate reproducibility. Appendices A and B describe the dataset construction pipeline, including video preprocessing, subtitle extraction, character library construction, speaker annotation, hierarchical video captioning, and QA generation and validation. They also provide prompts and annotation guidelines for LLM-assisted generation and filtering, human verification and editing, language prior removal, and global uniqueness checking. Appendices C and D detail the evaluation setup, including frame sampling, input preparation, model configurations, evaluation metrics, hardware, and inference-time measurement. Appendix E further describes the construction of multi-hop questions. The resulting dataset files and the code used for model evaluation are included in the supplementary materials to support independent verification and reproduction of our results.

REFERENCES

Mohamad Abou Ali, Fadi Dornaika, and Jinan Charafeddine. Agentic ai: a comprehensive survey of architectures, applications, and future directions. *Artificial Intelligence Review*, 59(1):11, 2025.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Brandon Castellano. Pyscenedetect: Video cut detection and analysis tool, 2024. URL `https://github.com/Breakthrough/PySceneDetect`.

Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. *arXiv preprint arXiv:2412.12075*, 2024a.

Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen, and Jiajun Qi. An enhanced res2net with local and global feature fusion for speaker verification. *arXiv preprint arXiv:2305.12838*, 2023.

Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Tinglong Zhu, Rongjie Huang, Chong Deng, Qian Chen, Shiliang Zhang, Wen Wang, et al. 3d-speaker-toolkit: An open-source toolkit for multimodal speaker verification and diarization. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Computer Vision and Pattern Recognition*, 2024b.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025.

Seongho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Seungchan Lee, Minsu Lee, and Byoung-Tak Zhang. Dramaqa: Character-centered video story understanding with hierarchical qa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1173–1181, 2021.

Rohan Choudhury, Guanglei Zhu, Sihan Liu, Koichiro Niinuma, Kris Kitani, and László Jeni. Don't look twice: Faster video transformers with run-length tokenization. *Advances in Neural Information Processing Systems*, 2024.

Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, et al. Paddleocr-vl: Boosting multilingual document parsing via a 0.9 b ultra-compact vision-language model. *arXiv preprint arXiv:2510.14528*, 2025a.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*, 2025b.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Computer Vision and Pattern Recognition*, 2019.

Anxhelo Diko, Tinghuai Wang, Wassim Swaileh, Shiyan Sun, and Ioannis Patras. Rewind: Understanding long videos with instructed learnable memory. In *Computer Vision and Pattern Recognition*, 2025.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, 2024.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Computer Vision and Pattern Recognition*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, et al. Memory in the age of ai agents. *arXiv preprint arXiv:2512.13564*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Hongbo Jin, Qingyuan Wang, Wenhao Zhang, Yang Liu, and Sijie Cheng. Videomem: Enhancing ultra-long video understanding via adaptive memory management. *arXiv preprint arXiv:2512.04540*, 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Seon-Ho Lee, Jue Wang, Zhikang Zhang, David Fan, and Xinyu Li. Video token merging for long-form video understanding. *arXiv preprint arXiv:2410.23782*, 2024.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *Empirical Methods in Natural Language Processing*, 2018.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024b.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Computer Vision and Pattern Recognition*, 2024c.

Xirui Li, Chao Ma, Xiaokang Yang, and Ming-Hsuan Yang. Vidtome: Video token merging for zero-shot video editing. In *Computer Vision and Pattern Recognition*, 2024d.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Empirical Methods in Natural Language Processing*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 2023.

Yuanxin Liu, Kun Ouyang, Haoning Wu, Yi Liu, Lin Sui, Xinhao Li, Yan Zhong, Y Charles, Xinyu Zhou, and Xu Sun. Videoreasonbench: Can mllms perform vision-centric complex video reasoning? *arXiv preprint arXiv:2505.23359*, 2025.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 2023.

OpenAI. Introducing gpt-5.5. `https://openai.com/index/introducing-gpt-5-5/`, 4 2026. Accessed: 2026-09-22.

Kun Ouyang, Yuanxin Liu, Linli Yao, Yishuo Cai, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Conan: Progressive learning to reason like a detective over multi-scale visual evidence. *arXiv preprint arXiv:2510.20470*, 2025.

Jihao Qiu, Lingxi Xie, Xinyue Huo, Qi Tian, and Qixiang Ye. Longvideo-r1: Smart navigation for low-cost long video understanding. *arXiv preprint arXiv:2602.20913*, 2026.

Qwen Team. Qwen3.6-Plus: Towards real world agents, April 2026. URL `https://qwen.ai/blog?id=qwen3.6`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Stephen Robertson and Hugo Zaragoza. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc, 2009.

Bytedance Seed. Seed2.0 model card: Towards intelligence frontier for real-world complexity, February 2026. URL `https://lf3-static.bytednsdoc.com/obj/eden-cn/lapzild-tss/ljhwZthlaukjlkulzlp/seed2/0214/Seed2.0%20Model%20Card.pdf`.

Leqi Shen, Tianxiang Hao, Tao He, Sicheng Zhao, Yifeng Zhang, Pengzhang Liu, Yongjun Bao, and Guiguang Ding. Tempme: Video temporal token merging for efficient text-video retrieval. *arXiv preprint arXiv:2409.01156*, 2024a.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024b.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Computer Vision and Pattern Recognition*, 2024.

Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. Moviechat+: Question-aware sparse memory for long video question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48:374–389, 2026.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn 2: Captioning-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025a.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Computer Vision and Pattern Recognition Conference*, 2025b.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Computer Vision and Pattern Recognition*, 2016.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Zhipu AI GLM Team. Glm-5.3: Frontier coding with emergent cyber capabilities. https://z.ai/blog/glm-5.3, 8 2026b. Accessed: 2026-09-21.

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. *arXiv preprint arXiv:2506.13654*, 2025.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri Iyer, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *International Conference on Computer Vision*, 2025a.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, 2024b.

Yuxuan Wang, Yiqi Song, Cihang Xie, Yang Liu, and Zilong Zheng. Videollamb: Long streaming video understanding with recurrent memory bridges. In *International Conference on Computer Vision*, 2025b.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Computer Vision and Pattern Recognition*, pages 3272–3283, 2025c.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 2024.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Huanjin Yao, Ruifei Zhang, Jiaxing Huang, Jingyi Zhang, Yibo Wang, Bo Fang, Ruolin Zhu, Yongcheng Jing, Shunyu Liu, Guanbin Li, et al. A survey on agentic multimodal large language models. *arXiv preprint arXiv:2510.10991*, 2025.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI Conference on Artificial Intelligence*, 2019.

Ce Zhang, Yan-Bo Lin, Ziyang Wang, Mohit Bansal, and Gedas Bertasius. Silvr: A simple language-based video reasoning framework. *arXiv preprint arXiv:2505.24869*, 2025a.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*, 2024a.

Hongjie Zhang, Lu Dong, Yi Liu, Yifei Huang, Yali Wang, Limin Wang, and Yu Qiao. Lvbench: A benchmark for long-form video understanding with versatile multi-modal question answering. *International Journal of Computer Vision*, 133(11):7726–7747, 2025b.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024b.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Orr Zohar, Xiaohan Wang, Rui Li, Andrés Marafioti, Miquel Farré, Merve Noyan, Leandro von Werra, Serena Yeung-Levy, and Thomas Wolf. Apollo2: Exploring the long-video frontier of large multimodal models, 2025. URL `https://huggingface.co/datasets/Apollo-LMMs/LongTimeScope`.

Jialong Zuo, Yongtai Deng, Lingdong Kong, Jingkang Yang, Rui Jin, Yiwei Zhang, Nong Sang, Liang Pan, Ziwei Liu, and Changxin Gao. Videolucy: Deep memory backtracking for long video understanding. *arXiv preprint arXiv:2510.12422*, 2025.