

SAGAQA: A Multi-hop Reasoning Benchmark for Long-form Narrative Understanding in TV Series

Galann Pennec^{∞,◇,♡} Zhengyuan Liu[◇]

Nicholas Asher[§] Philippe Muller[∞] Nancy F. Chen[◇]

[∞]IRIT, University of Toulouse, France

[◇]Agency for Science, Technology and Research (A*STAR), Singapore

[♡]CNRS@CREATE, Singapore [§]CNRS, IRIT, France

galann.pennec@cnrsatcreate.sg, {liu_zhengyuan,nancy_chen}@a-star.edu.sg

{nicholas.asher,philippe.muller}@irit.fr

Abstract

We introduce SAGAQA, a long-form video benchmark for multi-hop reasoning over full-length TV series. Existing video reasoning benchmarks often emphasize local understanding of adjacent frames or clips. SAGAQA addresses this gap by requiring high-level comprehension of extended multimodal narratives in entire TV shows. A distinguishing feature of SAGAQA is the granularity of its reasoning steps. Our dataset necessitates long-range reasoning hops to connect information across completely different episodes. This requires models to reason over entire events and actions, demanding a deep understanding of the show’s narration and progression at a multimodal level. Motivated by recent progress in agentic methods, we further study how different planning strategies handle such complex reasoning. We categorize these approaches into three classes—*Parallel*, *Sequential*, and *Hybrid* planners—and evaluate their ability to generate coherent and complete reasoning plans. Our results on SAGAQA suggest that hybrid planners consistently produce higher-quality plans and exhibit stronger capabilities for complex, high-level narrative understanding in TV shows.¹

1 Introduction

Recent advancements in Multimodal Large Language Models (MLLMs) have significantly expanded the frontiers of video understanding, with models demonstrating impressive capabilities in visual reasoning (Feng et al., 2025; Wang et al., 2025; Chen et al., 2025). However, a critical disparity remains between the ability to perform short-term visual perception and the capacity for long-term narrative comprehension. While existing benchmarks have successfully formalized stepwise causal reasoning within short videos of a few minutes (Cheng

et al., 2025; Li et al., 2025) or multi-step reasoning over longer single-file narratives (Yu et al., 2025a), they typically rely on adjacent clips and low-level understanding.

To bridge the gap between short-context perception and long-form understanding, recent methodologies have largely bifurcated into efficient retrieval and agentic reasoning. Retrieval-Augmented Generation (RAG) approaches (Ren et al., 2025; Xu et al., 2025; Mao et al., 2025) and frame selection strategies (Yu et al., 2025b; Ye et al., 2025; Hu et al., 2025) attempt to manage extreme token lengths by isolating key visual information. Simultaneously, agent-based frameworks (Wang et al., 2024b; Yang et al., 2024) and tool-augmented reinforcement learning methods (Zhang et al., 2025b; Yuan et al., 2025) have introduced interactive reasoning to navigate video content. Despite these innovations, evaluations remain constrained, as they often overlook the high-level narrative progression essential for comprehending complex, long-form content such as TV series.

In this paper, we present SAGAQA, an abstractive question answering benchmark designed to evaluate multi-hop narrative reasoning over full-length TV series. SAGAQA addresses key limitations of prior datasets by introducing a *multi-video*, *multi-hop* and *multi-modal* benchmark, where inputs are substantially longer and narratively more complex (see Table 1). Each question targets a window of 20 consecutive episodes—approximately 20 hours of video—requiring models to perform “big-hop” reasoning to connect distant events from the same narrative chain. On average, answering a question from our dataset involves four reasoning hops, with relevant events potentially separated by up to 20 episodes. Therefore, SAGAQA constitutes a particularly challenging benchmark, especially in comparison to prior work that does not require the same depth of reasoning over complex narrative

¹Our code and benchmark are available at <https://github.com/galannp/sagaqa>

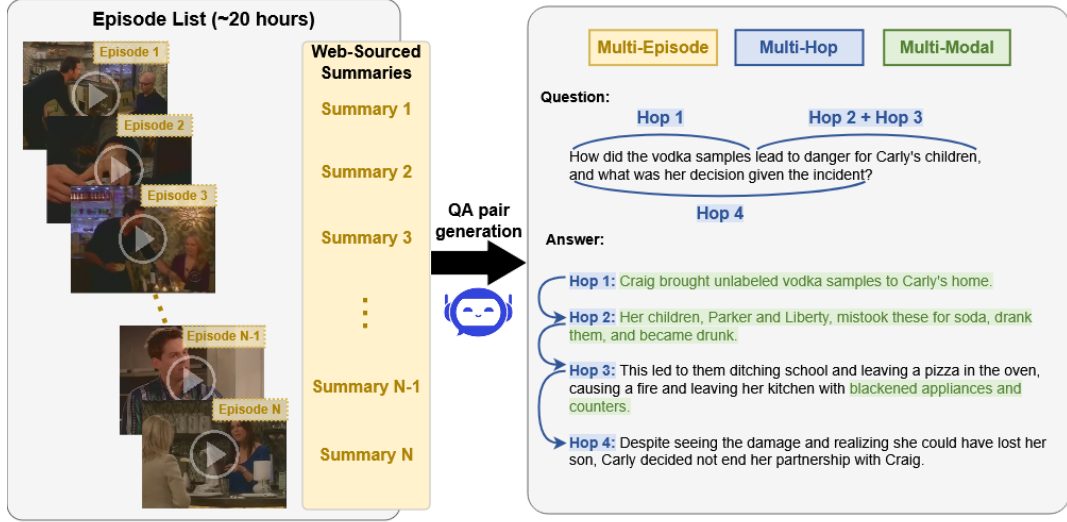


Figure 1: **Our data generation pipeline for SAGAQA.** 1) Given a list of 20 consecutive episodes $\mathcal{E}_{20}$ and their corresponding annotated summaries from SummScreen3D, we prompt an LLM to generate synthetic QA pairs that satisfy three key criteria: *multi-episode*, *multi-hop* and *multi-modal*. In the example shown, the generated question requires four reasoning hops, each being logically connected to the previous one in the reasoning chain. The hops are annotated in BLUE while the part of the answer that refers to both modalities (video and dialogue) is highlighted in GREEN.

video content (see Table 1)—such as egocentric video understanding (Yang et al., 2025a) or open-world internet video browsing (Liang et al., 2026).

The dataset is constructed using the pipeline in Figure 1. The resulting question-answer (QA) pairs undergo LLM-based filtering to ensure they satisfy the intended multi-hop reasoning criteria.

We study the performance of existing baselines on this benchmark including video planning methods, which we categorize into three classes: *Parallel*, *Sequential*, and *Hybrid*. Our results show that hybrid planners consistently outperform the other approaches, achieving more complete and accurate episode grounding by combining broad exploration of candidate video segments with focused reasoning over the most relevant ones. By moving evaluations toward long-term, event-driven comprehension, SAGAQA provides valuable insights into the current limitations and future directions of multi-step reasoning in long-form video understanding.

Our contributions are as follows:

- We introduce SAGAQA, a novel benchmark specifically designed for evaluating long-form, multi-hop reasoning over TV series, addressing the gap between short-term visual perception and full-episode narrative understanding.
- We systematically analyze video planning

strategies, introducing the distinction between *Parallel*, *Sequential*, and *Hybrid* Planners.

- We demonstrate that hybrid planners achieve superior performance by balancing efficient exploration with deep multi-hop reasoning, providing insights for future research in long-form video understanding.

2 Related Work

Long Video Understanding To evaluate the long-form understanding and reasoning capabilities of Video Large Language Models (Video-LLMs), several comprehensive benchmarks have been introduced (Mangalam et al., 2023; Wang et al., 2024a; Fu et al., 2024; Zhou et al., 2024; Li et al., 2024). These are categorized as *Long Video Understanding* datasets in Table 1. To tackle these benchmarks, existing methods typically rely on query-guided retrieval to extract relevant information (often specific frames) from the entire video. This is generally achieved through zero-shot extraction (Huang et al., 2025; Park et al., 2024; Wang et al., 2024c), tailored pretraining objectives (Yu et al., 2023, 2025b; Korbar et al., 2024), or agentic frameworks (Wang et al., 2024b; Yang et al., 2024).

Multi-Hop Reasoning and Agentic Approaches As video understanding has shifted from simple visual recognition to complex multi-hop reasoning,

Dataset	Domain	Duration	M-Vid	M-Hop	Hop Length	Reasoning	M-Mod	#Vid.
Long Video Understanding								
EgoSchema (Mangalam et al., 2023)	Egocentric	3min	✗	✗	—	—	✗	5,063
Video-MME (Fu et al., 2024)	Open Domain	1–60min	✗	✗	—	—	✓	900
LVBench (Wang et al., 2024a)	Open Domain	Up to 2h	✗	✗	—	—	✗	103
Multi-Hop Reasoning								
CausalStep (Li et al., 2025)	Open Domain	7min	✗	✓	sec-min	Narrative	✗	100
Video-Holmes (Cheng et al., 2025)	Short Films	1–5min	✗	✓	sec-min	Narrative	✓	270
VRBench (Yu et al., 2025a)	Open Domain	1.6h	✗	✓	Up to 40min+	Narrative	✓	960
Ultra-Long Video Reasoning								
Video-BrowseComp (Liang et al., 2026)	Open Web	hours	✓	✓	Search-Based	Fact-Finding	✓	Web
Ego-LifeQA (Yang et al., 2025a)	Egocentric	44.3h	✗	✓	Up to 24h+	Temporal	✓	6
SagaQA (ours)	TV Series	20h	✓	✓	Up to 20h	Narrative	✓	525

Table 1: **Comparison of SAGAQA with Existing Video Reasoning Benchmarks.** M-Vid: Multi-Video, M-Hop: Multi-Hop, M-Mod: Multi-Modal, #Vid.: Number of Videos.

models are now required to logically connect disparate temporal events (Yu et al., 2025a; Cheng et al., 2025; Li et al., 2025). The corresponding datasets are categorized as *Multi-Hop Reasoning* in Table 1. To bridge the gap between raw visual recognition and high-level understanding, Chain-of-Thought (CoT) methodologies have become essential (Han et al., 2025), alongside a growing line of agentic approaches (Zhi et al., 2025; Zhang et al., 2025b; Yuan et al., 2025; Zhang et al., 2025a). However, recent research highlights significant limitations in agentic video understanding. Autonomous models frequently suffer from *reasoning drift*, where their logic diverges from the original query over time, a problem often compounded by *noise contamination* from imperfect visual tools (Zhi et al., 2025; Liu et al., 2026). Motivated by the above limitations, our study compares and evaluates different agentic strategies under highly complex reasoning constraints from our SAGAQA dataset.

Reasoning on Ultra-Long Video Narratives Recent work on ultra-long video understanding has heavily focused on egocentric settings, where untrimmed recordings span from hours to entire days of human activity (Chandrasegaran et al., 2024; Yang et al., 2025a). Because end-to-end models struggle at this scale, recent methods introduce modular reasoning paradigms, such as Chain-of-Tool-Thought (Tian et al., 2025; Rege et al., 2026), to decouple high-level planning from tool execution. A parallel line of work explores similar agentic approaches for large-scale video browsing over the open web (Liang et al., 2026).

While these prior works push the temporal boundaries of video reasoning, Table 1 shows they

primarily emphasize *Fact-Finding* over the web or *Temporal* understanding such as tracking when a person last interacted with an object in egocentric videos, as seen in EgoLifeQA (Yang et al., 2025a). In contrast, the SAGAQA dataset differs fundamentally from those previous works by requiring models to uncover deep *Narrative* structures and causal dependencies that span multiple TV episodes (4.2 episodes on average). Because these *Narrative* dependencies are widely separated in time, they span substantially longer temporal windows than those reported in prior works (see Table 1), presenting a novel challenge for video agents.

3 Benchmark Construction

3.1 Dataset Description

Our dataset, SAGAQA, contains 200 QA pairs spanning 4 years of the TV show *As the World Turns* (2007-2010), providing a sufficient resource for controlled evaluation of multi-hop reasoning across extended video narratives. Each sample consists of a question Q associated with a set of 20 consecutive episodes $\mathcal{E}_{20}$, for which the corresponding videos are provided. The task is framed as abstractive question answering; the ground-truth annotations include the answer A to the question and the subset $\mathcal{E}^*$ of episodes that are relevant for answering it. Formally, a sample can be represented as $(Q, \mathcal{E}_{20}, A, \mathcal{E}^*)$.

We impose three key constraints on the QA pairs in the SAGAQA dataset:

- **Multi-Episode** The questions cannot be answered using a single episode.
- **Multi-Hop** The questions must integrate information from multiple audio-visual segments

through a series of dependent reasoning steps that progressively build toward the final answer.

- **Multi-Modal** The questions cannot be answered from dialogue alone and require analyzing visual elements from the video.

In Appendix A, we provide three dataset samples with manual annotations of the multi-hop reasoning path, the relevant episodes, and the multimodal elements that contribute to each answer.

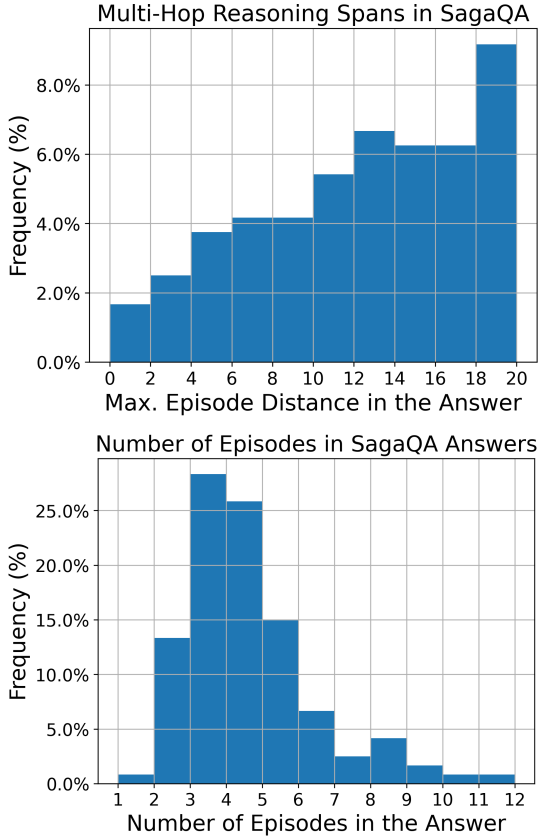


Figure 2: **Statistics of SAGAQA.** We provide the distribution of the maximum distance between two episodes involved in the multi-hop reasoning sequence as well as the distribution of the number of episodes involved in answering per question for all the questions in our dataset.

3.2 Dataset Statistics

On average, each question involves approximately 4.2 episodes and rarely exceeds 7 episodes, reflecting the high number of hops required for the questions (Figure 2).

The size of each hop, however, varies substantially within the 20-episode window. As illustrated by the distribution in Figure 2, most questions in our dataset exhibit relatively long depen-

dency chains, with a median temporal length of 13 episodes and a standard deviation of 5.4 episodes. Given that each episode has a fixed duration of one hour, the resulting reasoning chains span unusually long temporal intervals compared to other datasets in the literature (see Table 1).

3.3 QA Pairs Generation

We rely on existing annotations in the SummScreen3D dataset (Papalampidi and Lapata, 2023). SummScreen3D is a video-to-text summarization dataset of 5421 episodes each from popular soap operas like *As the World Turns*. It includes full transcripts, videos and multiple summaries for each episode. The summaries in SummScreen3D are also highly multimodal as they contain information referring to both the episode video and transcripts. This is crucial for generating our multimodal QA pairs.

For building our SAGAQA dataset we focus on the TV show *As the World Turns* as we notice that the TV show requires a proper multimodal understanding beyond simply the dialogue. Each QA pair in our dataset spans over 20 consecutive episodes, corresponding to about 20 hours of videos.

We detail below, and illustrate in Figure 1, the iterative process for generating each sample $(Q, \mathcal{E}_{20}, A, \mathcal{E}^*)$ in SAGAQA. All prompts are given in Appendix B.1.

- **Step 1:** Identify all the characters listed within the 20-episode context window $\mathcal{E}_{20}$. We denote as $\mathcal{C}$, the list of all extracted characters from the corresponding 20 episode summaries in SummScreen3D.
- **Step 2:** Given a character in $\mathcal{C}$ and the 20 episode summaries, we generate a QA pair for that character. The prompt given to Gemini 2.5 Flash (Comanici et al., 2025) precisely states the constraints on the generated pair: *multi-episode*, *multi-hop* and *multi-modal* (see Appendix B.1). We denote as $\mathcal{C}_{QA}$ the list of characters involved in the QA pair returned by the LLM.
- **Step 3:** We remove $\mathcal{C}_{QA}$ from the set of all the characters $\mathcal{C}$.
- **Step 4:** We repeat from **Step 2** until $\mathcal{C}$ is empty.
- **Step 5:** We finally post-process all generated QA pairs to retain only the ones that are

truly multi-hop. We use the prompt in Appendix B.2.

3.4 Human Verification

We validate our data generation process using two human evaluators: one co-author and one external researcher. Each evaluator is provided with the same set of 25 QA pairs.

To support the evaluation process, annotators are allowed to consult the output of an external LLM (Gemini 2.5 Flash), but only through a restricted set of prompts to ensure consistent and reliable responses. Specifically, for each episode E_i , evaluators may access the LLM-generated answer A_i to the question, where A_i is directly produced from the respective annotated summary of the episode for more robustness.

We provide the following guidelines to assess each QA pair with respect to our criteria:

- **Multi-Episode** Verify that no single episode-level answer A_i contains sufficient information to derive the final answer A .
- **Episode List** Identify the subset of episodes $\mathcal{E}_{\text{human}}$ required to construct the final answer A . This corresponds to selecting all episodes E_i whose associated answers A_i contribute to A . We then compute Precision, Recall, and F1 scores between the episode list $\mathcal{E}^*$ in SAGAQA and the annotated reference $\mathcal{E}_{\text{human}}$.
- **Multi-Hop** Assess whether the different elements required to answer question Q are logically connected and interdependent.

We report below the average results found by both annotators.

- **Multi-Episode** 95% of the answers in SAGAQA involves combining information from different episodes.
- **Episode List** The average Precision, Recall and F1 scores for the episode list annotation of each sample are 87%, 93% and 90% respectively. We obtain a Jaccard similarity of 86% between the episode lists identified by the two annotators.
- **Multi-Hop** The step for filtering multi-hop questions (Step 5 of our data generation, Section 3.3) achieved a precision of 82%, recall

of 91%, and an F1 score of 86%. Cohen’s κ was 0.61, indicating substantial agreement between the two annotators.

4 Multi-Hop Video Reasoning with Agentic Planners

4.1 Video Planning

Long TV shows are too large for end-to-end processing by VLMs. Additionally, multi-hop questions often require step-by-step reasoning, making static single-turn retrieval highly ineffective (Press et al., 2023). Multi-hop reasoning is commonly addressed through planning strategies in the literature. A planner aims to answer a composite question, i.e., a question that requires both independent (parallel) and dependent (multi-hop) reasoning steps to produce a final answer. Typically, it generates a set of queries that are executed by external tools and whose outputs are aggregated to construct a global answer. Video planners can be categorized into three types (Figure 3).

Parallel Planner A planner that decomposes a composite question into a set of simpler, independent subquestions, which are then answered in parallel (Rege et al., 2026).

Sequential Planner A planner that answers a composite question through a sequence of steps (Zhang et al., 2025b; Yuan et al., 2025; Wang et al., 2024b; Tian et al., 2025). At each step (Figure 3), the agent selects an action from a predefined set, generates a subquestion, and invokes the corresponding tools. Overall, the sequential planner issues queries one at a time, progressively building toward the final answer.

Hybrid Planner We introduce the hybrid planner as a way to combine the strengths of both parallel and sequential approaches. Like the sequential planner, the agent selects an action at each step. However, similar to the parallel planner, it generates and executes multiple independent queries simultaneously at each step. This design significantly reduces the length of reasoning chains. It mitigates the main limitations of sequential planning: 1) reasoning traces tend to drift and accumulate noise as depth increases, and 2) low cost efficiency due to the inability to execute queries in parallel. At the same time, it preserves the flexibility needed to handle multi-hop dependencies, which purely parallel planners fail to capture.

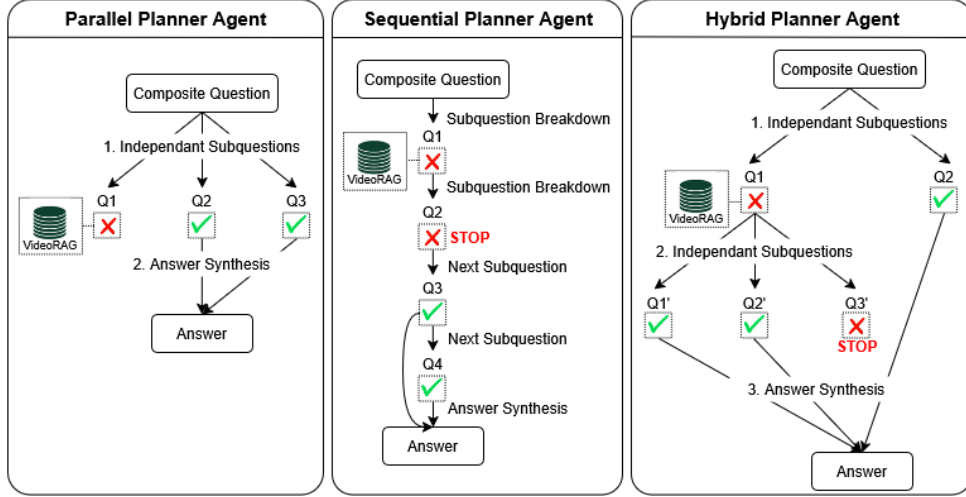


Figure 3: **Existing strategies for video planning.** From left to right: 1) the *Parallel Planner* splits a composite question into a set of independent subquestions; 2) the *Sequential Planner* generates subquestions one at a time, progressively working toward an answer; 3) the *Hybrid Planner* combines the two previous strategies by iteratively producing independent requests to the tools while eventually constructing an answer. **STOP** denotes the end of a reasoning trajectory executed by an agent. The green checkmark indicates that the tools successfully produced an answer, while the red cross indicates a failure.

4.2 Video Planners Implementation

In this section, we describe each studied video planning strategy and provide full implementation details for all methods in Appendix B.3. In our experiments, we use the VideoRAG framework (Ren et al., 2025) as the only tool for retrieving video information. In the VideoRAG implementation, an LLM judges the confidence of the retrieved video segments. If all retrieved segments are assigned low confidence, the system reports back to the planner that no answer was found for the query (✗).

Parallel Planner + VideoRAG We study two variants of this approach. The first implementation (Prompt 1) extracts keywords from the question, groups them by similarity, and generates one subquestion per group to query the VideoRAG. In the second implementation (Prompt 2), an LLM performs an end-to-end decomposition of the original composite question into subquestions.

Sequential Planner + VideoRAG In this setting, we choose the set of actions from among (BREAKDOWN, MOVE_ON).

The BREAKDOWN action is selected when the previous VideoRAG query fails to yield a satisfactory answer (✗), prompting the agent to further decompose the query.

The MOVE_ON action is selected either when the answer to the previous subquestion was satisfactory (✓) or when repeated attempts fail to produce useful results (**STOP**). In both cases, the agent gen-

erates the next subquestion, targeting a different aspect of the original query.

Hybrid Planner + VideoRAG Similar to the sequential planner, the agent dynamically selects between the BREAKDOWN and MOVE_ON actions. However, in this setting, the BREAKDOWN action is executed in parallel: instead of generating subquestions sequentially, the agent produces a set of independent subquestions that are queried concurrently to the VideoRAG.

4.3 Implementation Details

Unlike the original implementation of VideoRAG, which relies on Gemma-2-9B (Team, 2024), we use Qwen3-30B-Instruct (Team, 2025) or Mistral-Small-3.1-24B (Mistral AI, 2025) as the language model for both planning and final answer generation. In the prompt to all video planners, we always fix the length of the final answer to 200 words (see Appendix B.3.5). We limit the maximum depth of the reasoning sequence to 15 in the sequential planner as we notice that reasoning beyond that limit almost always ends up looping over the same reasoning steps. For the hybrid planner, we limit the depth of the reasoning tree to 2.

5 Experiments

5.1 Evaluation Metrics

Text Generation Metrics We report ROUGE-1 (r1), ROUGE-2 (r2), and ROUGE-Lsum (rlsum)

	r1	r2	rlsum	METEOR	Precision	Recall	F1
Non-Planning							
TextRAG	36.30	6.51	33.09	23.49	28.13	11.14	14.92
VideoRAG	36.71	6.72	33.27	24.42	36.91	30.46	28.54
Parallel Planners							
EGAgent (Qwen3-30B-Instruct)	33.01	4.25	29.41	20.38	38.89	33.33	32.86
<i>Parallel Planner (Prompt 1) + VideoRAG</i>	35.00	5.93	31.93	23.42	50.03	26.35	30.88
<i>Parallel Planner (Prompt 2) + VideoRAG</i>	36.06	6.15	32.98	24.10	48.06	30.93	34.26
Sequential Planners							
VideoExplorer (7B)	27.18	10.67	23.89	18.94	33.33	10.42	15.83
<i>Sequential Planner + VideoRAG</i>	36.94	6.67	33.35	24.22	46.76	30.43	32.68
Hybrid Planners							
<i>Hybrid Planner + VideoRAG</i>	35.99	6.32	32.58	24.07	45.47	44.11	41.02

Table 2: **Results on the test set with Qwen3-30B-Instruct.** We report ROUGE-1 (r1), ROUGE-2 (r2), ROUGE-Lsum (rlsum) and METEOR. We also include episode grounding metrics with the Precision, Recall and F1 scores. Best results are in **bold**.

using the python-rouge package, as well as METEOR scores computed with the `meteor_score` function from `nltk.translate`.

Episode Grounding Metrics We report Precision, Recall and F1 to assess if the episodes retrieved by the system match the annotations in SAGAQA.

5.2 Comparative Baselines

We group the baselines into planning and non-planning methods in Table 2. Among the planning methods, VideoExplorer is a trained sequential planner, while EGAgent follows a parallel planning strategy. We exclude agentic baselines that focus solely on visual information (Wang et al., 2024b; Yang et al., 2025c), as our dataset requires multimodal understanding of audio and visual signals. However, we include TextRAG as a single-modality baseline to specifically assess the role of multimodal reasoning in SAGAQA. Finally, given the extreme length of TV show content, direct end-to-end reasoning with Video-LLMs is computationally infeasible.

VideoRAG (Ren et al., 2025) Retrieval Augmented Generation (RAG) implementation for scalable understanding and generation over hours of videos along with audio.

TextRAG We modify the above VideoRAG implementation to perform query matching solely based on the automatically transcribed subtitles.

VideoExplorer (7B) (Yuan et al., 2025) VideoExplorer is an agentic long-video understanding

framework that is trained to iteratively plan sub-questions, performs temporal grounding, and conducts task-driven video exploration to enable scalable reasoning over extended video content. We let the agent process both the video and transcribed subtitles with Whisper-Small (Radford et al., 2023) provided in the whisper Python package.

EGAgent (with Qwen3-30B-Instruct) (Rege et al., 2026) EGAgent is a framework for long video understanding that leverages entity-centric scene graphs and hybrid multimodal retrieval for multi-hop reasoning.

5.3 Results

We split the dataset into two subsets, using 50 QA pairs for validation and 150 for testing. For each tested baseline, we identify the set of episodes used to ground its generated answers. This predicted episode set is then compared to the annotations in SAGAQA to compute the Precision, Recall, and F1 scores reported in Table 2. For methods relying on VideoRAG, this set corresponds to the episodes from which high-confidence video segments are retrieved; for VideoExplorer and EGAgent, we instead parse the model outputs to extract the episodes used to support their final answers.

The Importance of Multimodality We notice that all the episode grounding metrics drop when using TextRAG instead of VideoRAG, suggesting that visual information plays a meaningful role in answering the questions in SAGAQA.

Hybrid Planning for Better Episode Grounding

The hybrid planner achieves substantially stronger episode grounding than the other evaluated video planners as indicated by higher F1 scores in Table 2. This improvement indicates that it effectively combines the strengths of parallel and sequential strategies, leveraging both broad exploration and deep reasoning (see Section 6).

Importantly, these observations hold consistently when replacing Qwen3-30B-Instruct with Mistral-Small-3.1-24B in all VideoRAG-based methods (see Appendix C).

6 Discussion

Exploration–Exploitation Trade-Off Our experiments reveal a systematic trade-off across planners. Parallel planners issue multiple queries in a single step, enabling broad coverage of the search space but limiting their ability to perform iterative, multi-hop refinement. In contrast, sequential planners refine queries over multiple steps, but tend to over-exploit early high-confidence directions, repeatedly issuing similar queries and failing to adequately cover alternative directions. Our hybrid planner mitigates both limitations and achieves significantly higher F1 scores (Table 2). To illustrate this behavior, we compare, in Appendix D, the reasoning traces of our hybrid planner with those of two sequential planners, highlighting the latter’s tendency to repeatedly issue similar queries instead of effectively expanding the search space.

Limitations in Answer Synthesis Even when the planner retrieves the relevant video segments required to answer the question, effectively exploiting this evidence to produce a high-quality answer remains challenging. While episode grounding metrics vary significantly across the studied settings, abstractive metrics—such as ROUGE and METEOR—reported in Table 2 remain largely unchanged. This discrepancy suggests that current methods are limited not in retrieval, but in downstream information synthesis. In particular, even when the correct episode is identified, the planner may fail to capture the precise narrative pivot (e.g., a subtle facial expression or a key visual cue), since the agentic planner processes video segments as isolated units rather than as parts of a coherent temporal arc. This highlights the need for more structured post-retrieval reasoning and aggregation mechanisms.

Toward Structured and Hybrid Planning Strategies

The strong performance of the hybrid planner underscores the benefits of combining complementary reasoning paradigms. The parallel component acts as a broad exploration mechanism, identifying a diverse set of candidate video segments, while the sequential component serves as a knowledge refiner, ensuring that multi-hop dependencies are coherently connected. This division of roles suggests that future long-form agents should move beyond single-trajectory reasoning and instead adopt more structured exploration strategies, such as Tree-of-Thought (Yao et al., 2023), tree-based or graph-based planning frameworks (Yang et al., 2025b). Further improvements could be achieved through targeted training, as well as the integration of specialized retrieval and grounding tools to enhance both coverage and reasoning consistency.

7 Conclusion

In this paper, we introduced SAGAQA, a novel benchmark designed to push the boundaries of long-form video understanding from local perception to high-level narrative comprehension. By focusing on multi-hop reasoning across 20-episode windows—representing approximately 20 hours of continuous video—SAGAQA challenges models to connect disparate events, character arcs, and multimodal cues that span significant temporal distances.

Our systematic evaluation of various video planning strategies reveals a critical *exploration-exploitation* trade-off in long-form reasoning. While parallel planners provide broad coverage of the search space, they often lack the depth required for complex causal chains. Conversely, sequential planners excel at refinement but are prone to reasoning drift and repetitive loops. We demonstrate that hybrid planners, which combine parallel exploration and sequential refinement, achieve superior performance, particularly in grounding answers to the correct episodes.

Our results suggest that, beyond retrieval, a key bottleneck lies in effectively synthesizing evidence across extended narratives. Even when relevant episodes are correctly identified, current models struggle to consistently aggregate this information into accurate answers. By releasing SAGAQA and our evaluation framework, we aim to facilitate further research in long video reasoning.

Limitations

On the evaluation side, the proposed task-specific metrics—Episode Precision, Recall, and F1—are both challenging and sufficiently discriminative. However, the current annotations in SAGAQA could be further manually refined by experts by grounding the answers at the clip level rather than entire episodes. These finer-grained annotations would allow for a more precise and informative evaluation than episode-level assessment.

The benchmark is built exclusively from TV shows (Section 3.3). Soap operas have very specific narrative structures (e.g., strong cross-episode narrative continuity, dialogue-driven storytelling, and recurring characters). Models might overfit to this genre, and the results may not generalize to other video domains such as documentaries or instructional videos.

While 20 hours of video is long-form, the benchmark uses a fixed window of 20 consecutive episodes (Section 3.1). In actual TV series, character arcs and plot hops can span hundreds of episodes over several years. SAGAQA does not yet test *ultra-long* memory that spans the entire lifecycle of a TV series.

Acknowledgments

We thank the anonymous reviewers for their feedback. We are also grateful to Parth Sakhare for his assistance with the human evaluation of SAGAQA. This research is supported by the National Research Foundation, Prime Minister’s Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the National Research Foundation, Singapore.

References

- Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristóbal Eyza-guirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. 2024. [Hourvideo: 1-hour video-language understanding](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, Sifei Liu, Hongxu Yin, Yao Lu, and Song Han. 2025. [Scaling RL to long videos](#). *CoRR*, abs/2507.07966.
- Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. 2025. [Video-holmes: Can MLLM think like holmes for complex video reasoning?](#) *CoRR*, abs/2505.21374.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Naveen Sachdeva, Inderjit S. Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilai Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Sharon Silver, Ayzaan Wahid, Sergey Brin, Yves Raimond, Klemen Kloboves, Cindy Wang, Nitesh Bharadwaj Gundavarapu, Ilia Shumailov, Bo Wang, Mantas Pajarskas, Joe Heyward, Martin Nikolchev, Maciej Kula, Hao Zhou, Zachary Garrett, Sushant Kafle, Sercan Arik, Ankita Goel, Mingyao Yang, Jiho Park, Koji Kojima, Parsa Mahmoudieh, Koray Kavukcuoglu, Grace Chen, Doug Fritz, Anton Bulyenov, Sudeshna Roy, Dimitris Paparas, Hadar Shemtov, Bo-Juen Chen, Robin Strudel, David Reitter, Aurko Roy, Andrey Vlasov, Changwan Ryu, Chas Lechner, Haichuan Yang, Zelda Mariet, Denis Vnukov, Tim Sohn, Amy Stuart, Wei Liang, Minmin Chen, Praynaa Rawlani, Christy Koh, JD Co-Reyes, Guangda Lai, Praseem Banzal, Dimitrios Vytiniotis, Jieru Mei, and Mu Cai. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *CoRR*, abs/2507.06261.
- Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. 2025. [Video-R1: Reinforcing video reasoning in mllms](#). *CoRR*, abs/2503.21776.
- Chaoyou Fu, Yuhao Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiaowu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. 2024. [Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis](#). *CoRR*, abs/2405.21075.
- Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. 2025. [Videospresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*

- 2025, Nashville, TN, USA, June 11-15, 2025, pages 26181–26191. Computer Vision Foundation / IEEE.
- Jian Hu, Zixu Cheng, Chenyang Si, Wei Li, and Shao-gang Gong. 2025. [Cos: Chain-of-shot prompting for long video understanding](#). *CoRR*, abs/2502.06428.
- De-An Huang, Subhashree Radhakrishnan, Zhiding Yu, and Jan Kautz. 2025. [FRAG: frame selection augmented generation for long video and long document understanding](#). *CoRR*, abs/2504.17447.
- Bruno Korbar, Yongqin Xian, Alessio Tonioni, Andrew Zisserman, and Federico Tombari. 2024. [Text-conditioned resampler for long form video understanding](#). In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXVI*, volume 15144 of *Lecture Notes in Computer Science*, pages 271–288. Springer.
- Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. 2024. [Mvbench: A comprehensive multi-modal video understanding benchmark](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 22195–22206. IEEE.
- Xuchen Li, Xuzhao Li, Shiyu Hu, Kaiqi Huang, and Wentao Zhang. 2025. [CausalStep: A benchmark for explicit stepwise causal reasoning in videos](#). *CoRR*, abs/2507.16878.
- Zhengyang Liang, Yan Shu, Xiangrui Liu, Minghao Qin, Kaixin Liang, Nicu Sebe, Zheng Liu, and Lizi Liao. 2026. [Video-browser: Towards agentic open-web video browsing](#).
- Chengwen Liu, Xiaomin Yu, Zhuoyue Chang, Zhe Huang, Shuo Zhang, Heng Lian, Kunyi Wang, Rui Xu, Sen Hu, Jianheng Hou, Hao Peng, Chengwei Qin, Xiaobin Hu, Hong Peng, Ronghao Chen, and Huacan Wang. 2026. [Watching, reasoning, and searching: A video deep research benchmark on open web for agentic video reasoning](#). *CoRR*, abs/2601.06943.
- Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. [Egoschema: A diagnostic benchmark for very long-form video language understanding](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Mingyang Mao, Mariela M. Perez-Cabarcas, Utteja Kallakuri, Nicholas R. Waytowich, Xiaomin Lin, and Tinoosh Mohsenin. 2025. [Multi-rag: A multimodal retrieval-augmented generation system for adaptive video understanding](#). *CoRR*, abs/2505.23990.
- Mistral AI. 2025. Mistral small 3.1. <https://mistral.ai/news/mistral-small-3-1>. Accessed: June 3, 2026.
- Pinelopi Papalampidi and Mirella Lapata. 2023. [Hierarchical3D adapters for long video-to-text summarization](#). In *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1267–1290. Association for Computational Linguistics.
- Jongwoo Park, Kanchana Ranasinghe, Kumara Kahatapitiya, Wonjeong Ryoo, Donghyun Kim, and Michael S. Ryoo. 2024. [Too many frames, not all useful: Efficient strategies for long-form video QA](#). *CoRR*, abs/2406.09396.
- Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. [Measuring and narrowing the compositionality gap in language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Findings of ACL, pages 5687–5711. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. [Robust speech recognition via large-scale weak supervision](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 28492–28518. PMLR.
- Aniket Rege, Arka Sadhu, Yuliang Li, Kejie Li, Ramya Korlakai Vinayak, Yuning Chai, Yong Jae Lee, and Hyo Jin Kim. 2026. [Agentic very long video understanding](#). *CoRR*, abs/2601.18157.
- Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. 2025. [Videorag: Retrieval-augmented generation with extreme long-context videos](#). *CoRR*, abs/2502.01549.
- Gemma Team. 2024. [Gemma 2: Improving open language models at a practical size](#). *CoRR*, abs/2408.00118.
- Qwen Team. 2025. [Qwen3 technical report](#). *CoRR*, abs/2505.09388.
- Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkan Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. 2025. [Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning](#). *CoRR*, abs/2506.13654.
- Wei Han Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. 2024a. [LVBench: An extreme long video understanding benchmark](#). *CoRR*, abs/2406.08035.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang,

- Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yanan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingdong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. 2025. [InternV3.5: Advancing open-source multi-modal models in versatility, reasoning, and efficiency](#). *CoRR*, abs/2508.18265.
- Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024b. [VideoAgent: Long-form video understanding with large language model as agent](#). In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXX*, volume 15138 of *Lecture Notes in Computer Science*, pages 58–76. Springer.
- Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2024c. [VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos](#). *CoRR*, abs/2405.19209.
- Zeyu Xu, Junkang Zhang, Qiang Wang, and Yi Liu. 2025. [E-VRAG: enhancing long video understanding with resource-efficient retrieval augmented generation](#). *CoRR*, abs/2508.01546.
- Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhonggang Cai, Bo Li, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, and Ziwei Liu. 2025a. [Egolife: Towards egocentric life assistant](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 28885–28900. Computer Vision Foundation / IEEE.
- Te Yang, Xiangyu Zhu, Bo Wang, Quan Chen, Peng Jiang, and Zhen Lei. 2025b. [EEA: exploration-exploitation agent for long video understanding](#). *CoRR*, abs/2512.03500.
- Zeyuan Yang, Delin Chen, Xueyang Yu, Maohao Shen, and Chuang Gan. 2024. [VCA: video curious agent for long video understanding](#). *CoRR*, abs/2412.10471.
- Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, and Lidong Bing. 2025c. [Longvt: Incentivizing "thinking with long videos" via native tool calling](#). *CoRR*, abs/2511.20785.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. [Tree of thoughts: Deliberate problem solving with large language models](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristóbal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, Jiajun Wu, and Manling Li. 2025. [Re-thinking temporal search for long-form video understanding](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 8579–8591. Computer Vision Foundation / IEEE.
- Jiashuo Yu, Yue Wu, Meng Chu, Zhifei Ren, Zizheng Huang, Pei Chu, Ruijie Zhang, Yanan He, Qirui Li, Songze Li, Zhenxiang Li, Zhongying Tu, Conghui He, Yu Qiao, Yali Wang, Yi Wang, and Limin Wang. 2025a. [VRBench: A benchmark for multi-step reasoning in long narrative videos](#). *CoRR*, abs/2506.10857.
- Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2023. [Self-chained image-language model for video localization and question answering](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, Hao Zhang, and Qianru Sun. 2025b. [Frame-Voyager: Learning to query frames for video large language models](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Ji-Rong Wen, and Zhicheng Dou. 2025. [Videodeep-research: Long video understanding with agentic tool using](#). *CoRR*, abs/2506.10821.
- Congzhi Zhang, Zhibin Wang, Yinchao Ma, Jiawei Peng, Yihan Wang, Qiang Zhou, Jun Song, and Bo Zheng. 2025a. [Rewatch-r1: Boosting complex video reasoning in large vision-language models through agentic data synthesis](#). *CoRR*, abs/2509.23652.
- Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. 2025b. [Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning](#). *CoRR*, abs/2508.04416.
- Zhuo Zhi, Qiangqiang Wu, Minghe shen, Wenbo Li, Yinchuan Li, Kun Shao, and Kaiwen Zhou. 2025. [VideoAgent2: Enhancing the LLM-based agent system for long-form video understanding by uncertainty-aware CoT](#). *CoRR*, abs/2504.04471.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. [MLVU: A comprehensive benchmark for multi-task long video understanding](#). *CoRR*, abs/2406.04264.

A SAGAQA Examples

We provide three different examples from SAGAQA. We also manually studied whether these samples satisfy the criteria defined for the dataset, namely *multi-episode*, *multi-hop* and *multi-modal*.

- **Multi-Episode** The answers explicitly refer to multiple episodes. We highlight in **YELLOW** references to the episodes.
- **Multi-Hop** We manually annotate in **BLUE** the reasoning hops required to answer each of the questions below. The questions require 5, 4 and 3 hops respectively.
- **Multi-Modal** After watching the episodes relevant to each of the questions, we manually highlight in **GREEN** the parts of the answer that directly refers to the video content.

A.1 Example 1

Question Q:

(Hop 1) How did Midnight Sun vodka samples (Hop 2) at Carly’s home (Hop 3 + Hop 4) lead to danger for her children, (Hop 5) and what was her decision about the business, given her past with an alcoholic father and the observable proof of the incident?

Episodes to Answer $\mathcal{E}^*$: 9th, 20th, February and 4th March 2009

Answer A:

(Hop 1) Carly, initially excited about her Midnight Sun vodka venture with Craig, (Hop 2) found herself in a dangerous situation when Craig brought unlabeled product samples to her home. (Hop 3) Her children, Parker and Liberty, mistook these for soda, drank them, and became drunk. (Hop 4) This led to them ditching school and leaving a pizza in the oven, causing a fire and leaving her kitchen with blackened appliances and counters. (Hop 5) Upon seeing the damage and realizing she could have lost her son, Carly was deeply

shocked and emotional. She revealed her childhood trauma of being terrorized by an alcoholic father, expressing her fear that Parker might follow a similar path. Despite this profound realization and the incident’s severity, Carly initially decided not to shelve the Midnight Sun project, affirming her partnership with Craig. However, the event profoundly underscored her deep-seated fears about alcohol’s impact on her family.

A.2 Example 2

Question Q: (Hop 1) After Emily hired her to separate Casey and Alison, (Hop 2 + Hop 3) how did Jade romantically pursue Casey, (Hop 4) and how did he outwardly respond to her advances?

Episodes to Answer $\mathcal{E}^*$: 6th and 19th February 2009

Answer A:

(Hop 1) On February 6th, after being secretly hired by Emily to keep Alison and Casey apart, (Hop 2) Jade developed her own romantic interest in Casey.

(Hop 3) On February 19th, after Casey cancelled their date, Jade, wearing her new dress, went to Memorial Hospital. She found Casey in an exam room, where she pulled the blinds, locked the door, and gave Casey an inviting kiss. (Hop 4) However, Casey pushed her away and made a break for the hall, clearly rejecting her advance.

A.3 Example 3

Question Q:

(Hop 1 + Hop 2) How did Craig create a secret bond with Parker, (Hop 3) and how did it lead to Craig secretly funding Parker’s Metallica tickets?

Episodes to Answer $\mathcal{E}^*$: 2nd, 9th and 30th January 2009

Answer A:

(Hop 1) On January 2nd, Craig gave Parker

his car keys for Parker’s 16th birthday, allowing him to drive to hockey practice. Parker, however, took Liberty on a joyride, and they ended up stranded and making out. Jack was furious at Craig for loaning the car.

(Hop 2) A week later, on January 9th, Parker and Liberty were caught by Craig making out behind the bar at Metro, where Parker was doing chores for Carly. Craig, still possessing a key, found them but covered for their presence, rushing Liberty out the back door before Carly arrived. This created a secret bond and leverage for Craig.

(Hop 3) Later, on January 30th, Parker needed \$200 for Metallica concert tickets, but Jack, facing financial strain, couldn’t provide the money and reiterated his disapproval of Craig’s influence. Craig, aware of Parker’s desire for the tickets and having previously covered for him, gave Parker money to buy toys for Johnny, explicitly telling him to keep extra for his trouble. Parker’s eyes lit up, and he used this money to buy the Metallica tickets, effectively having Craig fund his entertainment despite Jack’s strong objections.

B Prompts

B.1 QA Pair Generation

B.1.1 Character Identification Prompt

We provide the prompt below to identify all the characters mentioned in a single episode summary within the 20-episode context window $\mathcal{E}_{20}$. The prompt is given to Qwen3-30B-Instruct.

- **<SUMMARY_i>** is the groundtruth summary of episode E_i from the episode list $\mathcal{E}_{20}$.

<SUMMARY_i>

List the main characters actively participating in the above episode separated by a comma. I want only the characters who have a known first name.

Characters:

B.1.2 QA Pair Generation Prompt

We provide the prompt below to generate the QA pairs in SAGAQA. The prompt is given to Gemini-2.5-Flash.

- **DATE_EPISODE_i** is the date of the episode E_i .
- **<SUMMARY_i>** is the groundtruth summary of episode E_i .
- **CHARACTER** is the character from the TV show around which the question must be built.

Summary **DATE_EPISODE_1**:

<SUMMARY_1>

Summary **DATE_EPISODE_2**:

<SUMMARY_2>

...

Summary **DATE_EPISODE_20**:

<SUMMARY_20>

Answer in the following way:

1. Multimodal Multi-Episode Multi-Hop Question around character **<CHARACTER>** involving multiple visual elements or actions (Do not provide the answer within the question, Keep your question clear and simple):
2. Compact Question (Write the same question in about 30 words):
3. Justification why the question is Multi-Hop. For each hop, explain how it relates to the previous hops and which reasoning type is involved (Bridge Entities (Compositional), Comparison & Intersection, Causal & Counterfactual, Temporal Reasoning, Multi-Step Quantitative Reasoning):
 - (a) First Hop:
 - (b) Second Hop (Justify how it relies on previous hops and by which reasoning type): ...
4. Justification why the question is Multimodal
5. Justification why the question is Multi-Episode:
6. List of Episode involved (separated by comma). Example 6 January 2003, ...:

7. Answer to the above question in 200 words maximum:
8. List of Characters involved in the question (separated by comma):

B.1.3 Example Output

1. Multimodal Multi-Episode Multi-Hop Question: Considering Carly's initial excitement about the "Midnight Sun" vodka venture with Craig, how did the visual presence of the product samples at her home lead to a dangerous situation for her children, and what was her subsequent emotional realization and decision regarding the business, particularly in light of her past experiences with an alcoholic father, as evidenced by the blackened kitchen appliances?
2. Compact Question: How did "Midnight Sun" vodka samples at Carly's home lead to danger for her children, and what was her decision about the business, given her past with an alcoholic father and the visual evidence of the incident?
3. Justification why the question is Multi-Hop:
 - (a) First Hop (Bridge Entities): Identify Carly's involvement in the "Midnight Sun" vodka business with Craig. This establishes the core entity and her initial engagement.
 - (b) Second Hop (Causal & Counterfactual): Connect the vodka samples (a specific product of the business) being present at her home to her children, Parker and Liberty, consuming them and getting drunk. This links the business to a direct consequence involving her family.
 - (c) Third Hop (Causal & Counterfactual / Temporal Reasoning): Link the children's consumption and subsequent actions (leaving pizza in the

oven) to the fire and the visually evident "blackened kitchen appliances and counters," which then triggers Carly's emotional realization of the danger.

- (d) Fourth Hop (Causal & Counterfactual): Explain Carly's decision regarding the "Midnight Sun" business, specifically in light of her past trauma with an alcoholic father and the recent dangerous incident involving her children.
4. Justification why the question is Multimodal: The question explicitly refers to "visual presence of the product samples" (the vodka cans/bottles) and "blackened kitchen appliances and counters" as evidence of the incident and its impact on Carly. The act of "drinking" is also an action that would be visually depicted.
5. Justification why the question is Multi-Episode: The events described span multiple episodes: the inception of the "Midnight Sun" vodka business (09 February), the children consuming the samples and Carly's initial realization (20 February), and the subsequent fire and Carly's renewed emotional distress upon seeing the damaged kitchen (04 March).
6. List of Episode involved: 09 February 2009, 20 February 2009, 04 March 2009
7. Answer to the above question in 200 words maximum: Carly, initially excited about her "Midnight Sun" vodka venture with Craig, found herself in a dangerous situation when Craig brought unlabeled product samples to her home. Her children, Parker and Liberty, mistook these for soda, drank them, and became drunk. This led to them ditching school and leaving a pizza in the oven, causing a fire and leaving her kitchen with "blackened appliances and counters."

Upon seeing the damage and realizing she could have lost her son, Carly was deeply shocked and emotional. She revealed her childhood trauma of being ter-

rorized by an alcoholic father, expressing her fear that Parker might follow a similar path. Despite this profound realization and the incident's severity, Carly initially decided not to shelve the "Midnight Sun" project, affirming her partnership with Craig. However, the event profoundly underscored her deep-seated fears about alcohol's impact on her family.

8. List of Characters involved in the question: Carly Tenney, Craig Montgomery, Parker Snyder, Liberty Snyder, Johnny Donovan, Jack Snyder, Janet Snyder

B.2 Multi-Hop Filtering

We provide the prompt below to filter our generated QA pairs, keeping only the questions that are truly multi-hop. The prompt is given to Qwen3-30B-thinking.

- **<QUESTION>** is the question to classify as multi-hop or not.

Justify whether this question is truly multi-hop. you assume you have access to a document containing all the answers. No external knowledge is required.

Definition of Multi-Hop:

1. A question is multi-hop if it requires reasoning across at least two pieces of information to form a complete answer.
2. Requirements
 - (a) Inference required: Answers cannot be copied verbatim; they demand logical connection.
 - (b) The Hops can be part of the same narrative

You can compare the pros and cons and provide your final justification and answer.

Question: **<QUESTION>**

B.3 Video Planners Prompts

We provide the prompts below for the video planners we tested on SAGAQA. The prompts are given

to either Qwen3-30B-instruct or Mistral-Small-3.1-24B.

B.3.1 Parallel Planner (Prompt 1)

- **<QUESTION>** is a composite question about the TV show.

a) Keywords Extraction

We start by extracting all the keywords from the question to answer about the TV show.

Question: **<QUESTION>**

Decompose the question into keywords to the video. Keep only the most relevant keywords. You may drop non useful keywords that does not add to the meaning.

Keywords List:

1. ...
2. ...
- ...

b) Keywords Ordering

Given the list of keywords we previously extracted, we can reorganize them by difficulty.

Order the keywords from the easiest to match to a unique scene or video moment to the least easy to match:

Ordered Keywords List:

1. ...
2. ...
- ...

c) Keywords Grouping

We group the keywords that are related together.

Group the keywords together. Do not miss any keyword.

Grouped keywords:

**** Group 1: ... ****

1. ...
2. ...
- ...

**** Group 2: ... ****

1. ...
2. ...
- ...

d) Subquestion Generation for each Group

We generate a title for each group of keywords. This title is later used as a query to the VideoRAG.

Propose a title for each group in a few words. Always include the group keywords into the title. Always refer to the name of the characters in your title.

Group List:

**** Group 1 Title: ... ****

**** Group 2 Title: ... ****

...

B.3.2 Parallel Planner (Prompt 2)

- **<QUESTION>** is a composite question about the TV show.

Task Definition for the Querying Agent (Planner)

You are a Video Query Planner Agent. Your role is to retrieve information from a video in order to answer a complex, multi-hop question.

Your task is not to answer the question directly. Instead, you must plan and issue a sequence queries to the video.

Objective

Given a Question, you must decompose it into

simple video queries that, when answered sequentially, will provide all necessary evidence to fully answer the question.

Querying Rules

1. A query is always grounded into the question. Do not make up any new information not already present within the question.
2. All queries should target a distinct information (non-overlapping).
3. Regroup similar queries together. Do not over decompose the queries in too specific.
4. Queries should be very short and clear.
5. Rank your queries from easier to harder

Output Format

Do not include answers, summaries, or explanations — only the planned queries. Your output must be in the same format as in the examples below

Example:

Question: What visual actions directly resulted in the definitive exposure of Meg's deception about taking her medication, and who was the primary witness to this final revelation?

Queries:

- * Query 1: Meg's deception about taking her medication.
- * Query 2: Visual actions that exposed Meg's deception about taking her medication.
- * Query 3: The primary witness to the revelation of Meg's deception about taking her medication.

Question:

<QUESTION>

Queries:

- * Query 1: ...
- * Query 2: ...
- ...

B.3.3 Sequential Planner

At any given step i , we proceed as follows:

- Generate the subquestion for the i^{th} iteration.
- Answer subquestion i by querying the VideoRAG.

Note that we do not need to produce a subquestion for the first iteration, since the first subquestion is simply the main question.

Below, we provide the prompt given to the sequential planner agent at step i . As stated in Section 4.3, we limit the maximum depth of the reasoning sequence to 15 as we observe that reasoning beyond this limit does not yield further improvements.

- **<QUESTION>** is a composite question about the TV show.
- **<SUBQUESTION_ i >** is the subquestion generated by the sequential planner at step i .
- **<ANSWER_ i >** is the answer produced by the VideoRAG to the i^{th} subquestion.

Query 1: **<QUESTION>**

Answer to Query 1: **<ANSWER_1>**

Query 2: **<SUBQUESTION_2>**

Answer to Query 2: **<ANSWER_2>**

...

Query $i - 1$: **<SUBQUESTION_ $i - 1$ >**

Answer to Query $i - 1$: **<ANSWER_ $i - 1$ >**

You are a Video Query Planner Agent. Your goal is to answer Query 1.

Your task is not to answer Query 1 directly. Instead, you must plan and issue a sequence of queries to eventually answer Query 1.

Given the answers to the previous queries, you can choose between different actions:

1. **BREAK DOWN** the last query (e.g. I need to break down the last query into subqueries as the last query is too complex).
2. **MOVE_ON** to the next query (e.g. the answer to the last query is satisfying and

we can move on to the next query or we decide to move on after too many unsuccessful attempts to answer the last query).

Justify which action you want to perform. Given your action, what is the next query. Produce the next query only.

Querying Rules

1. Altogether, the queries should eventually help you to build your answer to Query 1.
2. A query is always grounded into the question. Do not make up any new information not already present within the question.
3. The Queries should be very short and clear.
4. Multiple **BREAK DOWNS** can be performed as long as they are justified.

Answer in the following way:

Justification: ...

Action performed: ...

Query i : ...

B.3.4 Hybrid Planner

The hybrid planner is a combination of both the parallel and sequential planners. We leverage the prompts of both methods in the implementation of the hybrid approach.

At any given step i , we proceed as follows:

- Prompt the agent for the next action (**BREAKDOWN** or **MOVE_ON**), using the same prompt as the sequential planner in Appendix B.3.3
- If the action is **BREAKDOWN**, we use the prompt from the parallel planner to decompose the current query into a list of independent subquestions (Appendix B.3.2).
- If the action is **MOVE_ON**, the agent processes the next available subquestion from the most recent **BREAKDOWN**. If no pending subquestions

remain across all prior BREAKDOWN steps, the reasoning process stops.

As specified in Section 4.3, we limit the maximum depth of the reasoning tree to 2 and do not allow any reasoning beyond that fixed limit.

B.3.5 Answer Synthesis Prompt

We provide below the prompt used by the video planners to generate their final answers. We ask the LLM to produce an answer in 200 words.

- **<QUESTION>** is a composite question about the TV show.
- **<SUBQUESTION_*i*>** is the subquestion generated by the video planner at the *i*th step.
- **<ANSWER_*i*>** is the answer produced by the VideoRAG to the *i*th subquestion.

Query: **<SUBQUESTION_1>**
 Answer in 200 words maximum: **<ANSWER_1>**
 Query: **<SUBQUESTION_2>**
 Answer in 200 words maximum: **<ANSWER_2>**
 ...
 Query: **<QUESTION>**
 Answer in 200 words maximum: ...

C Additional Experiments

In Table 3, we report the results using Mistral-Small-3.1-24B as both the planner and question-answering models. Our conclusions remain the same as in Section 5.3 where we used Qwen3-30B-Instruct (see Table 2)

D Reasoning Trace Examples

We compare the reasoning traces of the hybrid planner with those of two sequential planners on a question sampled from our dataset.

Composite Question: After Alison confronted Mick in the belfry about his actions towards Dr. Cantor, what visual action did Mick perform that indicated his guilt, and how did this specific visual action later manifest in Alison’s disturbed dream on January 13th?

The reasoning traces of the two sequential methods below illustrate how sequential planning can easily lead to repetitive queries, causing the agent to revisit similar queries instead of effectively covering all aspects of a multi-hop question. We use the same color to underlight semantically similar queries, illustrating the agent’s failure to sufficiently explore new queries.

D.0.1 VideoExplorer (Sequential Planner)

Q1: Locate the segment where Alison confronts Mick in the belfry about his actions towards Dr. Cantor. After this confrontation, what visual action did Mick perform that indicated his guilt?

Q2: Locate the segment where Mick is in the belfry and related to Dr. Cantor. After this segment, what visual action did Mick perform that indicated his guilt?

Q3: Locate the segment where Mick is in the belfry. After this segment, what visual action did Mick perform that indicated his guilt?

Q4: Locate the segment where Alison has a dream on January 13th. Describe the content of the dream, especially any visual actions by Mick that might indicate guilt.

Q5: Locate the segment where Mick appears. Describe his actions or context.

Q6: Locate the segment where Alison appears. Describe her actions or context.

Q7: Locate segments depicting dreams, nightmares, or similar sleep-related scenarios, particularly in a belfry setting. Describe the visual actions of characters within these dreams or nightmares, if any characters resemble Mick, and any visual actions that might indicate guilt.

Q8: Locate the segment where Alison appears. Describe her actions or context.

Q9: Locate the segment where Alison confronts Mick in a forested area. After this confrontation, what visual action does Mick

	r1	r2	rlsum	METEOR	Precision	Recall	F1
Non-Planning							
TextRAG	36.02	6.91	32.87	24.02	29.17	11.14	15.06
VideoRAG	37.64	7.56	34.19	24.05	37.95	30.04	28.30
Parallel Planners							
<i>Parallel Planner (Prompt 1) + VideoRAG</i>	36.74	7.41	33.29	25.04	38.11	26.54	27.39
<i>Parallel Planner (Prompt 2) + VideoRAG</i>	36.51	7.34	33.06	25.00	56.62	34.14	37.74
Sequential Planners							
<i>Sequential Planner + VideoRAG</i>	34.70	7.26	31.54	23.00	48.82	31.87	34.21
Hybrid Planners							
<i>Hybrid Planner + VideoRAG</i>	34.14	6.80	30.98	24.01	48.80	45.70	41.69

Table 3: **Results on the test set with Mistral-Small-3.1-24B.** Column descriptions are the same as in Table 2. Best results are in **bold**.

perform that indicates his guilt?

Q10: Is Alison confronting Mick in this segment? If so, what visual action does Mick perform that indicates his guilt?

D.0.2 Sequential Planner + VideoRAG

Q1: The visual action Mick performed after Alison confronted him in the belfry about his actions towards Dr. Cantor, and how that action manifested in Alison’s disturbed dream on January 13th.

ACTION: BREAK DOWN

Q2: What specific visual action did Mick perform in the belfry after Alison confronted him about Dr. Cantor?

ACTION: BREAK DOWN

Q3: What specific gesture or movement did Mick make immediately after Alison confronted him in the belfry?

ACTION: BREAK DOWN

Q4: What visual detail or action occurred in the belfry immediately after the confrontation between Alison and Mick?

ACTION: BREAK DOWN

Q5: What specific visual element in the belfry (e.g., bell, stone arch, shadow) was present during the confrontation between

Alison and Mick?

ACTION: BREAK DOWN

Q6: What distinctive background feature (e.g., large bell, cracked stone wall, hanging rope) was visible in the belfry during the confrontation?

ACTION: BREAK DOWN

Q7: What camera angle or movement (e.g., close-up, slow pan, overhead shot) was used during the scene in the belfry?

ACTION: MOVE ON

Q8: What specific visual element appeared in Alison’s dream on January 13th that could be linked to Mick’s actions in the belfry?

ACTION: BREAK DOWN

Q9: What object or image repeatedly appeared in Alison’s dream on January 13th?

ACTION: BREAK DOWN

Q10: What single recurring object (e.g., bell, rope, shadow, hand gesture) appeared in Alison’s dream on January 13th?

ACTION: MOVE ON

Q11: What specific physical gesture or movement did Mick make in the belfry that could indicate guilt (e.g., covering his face, looking away, trembling hand)?

ACTION: BREAK DOWN

Q12: Did Mick touch or interact with any object in the belfry immediately after the confrontation?

ACTION: BREAK DOWN

Q13: Did Mick touch the bell in the belfry immediately after the confrontation with Alison?

D.0.3 Hybrid Planner

Q1: The events involving Alison confronting Mick in the belfry about his actions towards Dr. Cantor.

ACTION: BREAKDOWN

Q1': The events that took place in the belfry, specifically involving Alison and Mick.

ACTION: MOVE ON

Q2': Mick's actions towards Dr. Cantor.

ACTION: MOVE ON

Q3': The location of the confrontation between Alison and Mick.

ACTION: MOVE ON

Q2: Alison's reasons for confronting Mick about his actions towards Dr. Cantor.

ACTION: BREAKDOWN

Q1'': The visual action Mick performed in the belfry that indicated his guilt.

ACTION: MOVE ON

Q2'': Mick's actions in the belfry.

ACTION: MOVE ON

Q3'': Visual actions performed by Mick in the belfry that suggest guilt.

ACTION: BREAKDOWN

Q3: How the visual action Mick performed in the belfry manifested in Alison's dream on January 13th.

ACTION: MOVE ON

Q1''': Mick's visual actions in the belfry.

ACTION: MOVE ON

Q2''': Alison's dream on January 13th.

ACTION: MOVE ON

Q3''': How Mick's visual actions in the belfry manifested in Alison's dream on January 13th.