Keep It Simple: Multi-Key Episodic Memory Retrieval for Ultra-Long Video Understanding

Yeeun Choi[1], Youngbeom Yoo[1], Joon-Young Lee[2], Hyolim Kang[1,†], and Seon Joo Kim[1,†]

[1] Yonsei University, Seoul, Republic of Korea
[2] Adobe Research, San Jose, CA, USA

Project Page: `https://choi-yeeun.github.io/MERIT`

Abstract. When videos extend from hours to days, directly processing them end-to-end becomes impractical for current Multi-modal Large Language Models (MLLMs). This ultra-long setting necessitates a two-stage paradigm: query-agnostic memory construction followed by retrieval-based inference. Prior work invests in complex memory construction to pre-model high-level relations in videos, despite not knowing the downstream query at build time. We instead prioritize high-recall retrievability during memory building, and defer query-specific, high-level relation composition to inference time. To this end, we propose **MERIT** (**M**ulti-key **E**pisodic **R**etrieval with **I**nference-time **T**emporal expansion), a simple yet effective agentic framework for ultra-long video understanding. First, we formulate an episodic multi-key representation that enables precise retrieval of fine-grained memories through a simple key-matching mechanism. Second, we introduce a neighbor filtering mechanism to capture broader semantic context without the massive computational overhead of global memory construction. This is achieved by expanding the temporal scope exclusively around the retrieved segments at inference time. By leveraging simple key-matching with this on-demand temporal expansion, MERIT achieves state-of-the-art performance across three long-video benchmarks: EgoLifeQA, LVBench, and Video-MME(Long).

1 Introduction

The rapid evolution of Multi-modal Large Language Models (MLLMs) [3, 4, 6, 20, 38] has shifted video understanding from short temporal clips to ultra-long streams spanning hours or even days. This shift enables the development of personal AI assistants capable of agentic reasoning over continuous egocentric or surveillance video. However, as temporal length scales dramatically, direct end-to-end modeling becomes infeasible, making external memory and Retrieval-Augmented Generation (RAG) [7, 9, 14] frameworks essential.

A defining property of ultra-long video understanding is the structural separation between memory construction time and inference time. During memory construction, the system processes the entire video without knowledge of future queries and builds an external memory. At inference time, a specific question is

[†]Co-corresponding authors.

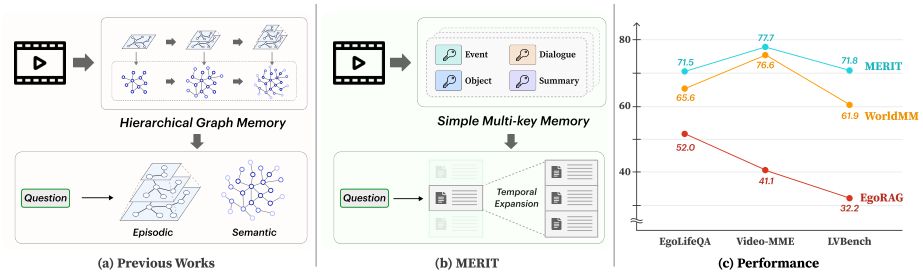

Fig. 1: (a) Previous methods: Construct hierarchical graph-based episodic and semantic memories, and at inference time, search and aggregate relevant information across these structured memories. (b) Our method: MERIT constructs a simple episodic multi-key memory, and at inference time, retrieves the matched clips along with temporal expansion to perform query-driven semantic reasoning. (c) Performance of MERIT against EgoRAG [31] and WorldMM [33] across various benchmarks [8, 27, 31].

posed, and the system must retrieve relevant evidence and perform reasoning. This decoupling fundamentally distinguishes long-video settings from conventional short-video Question Answering (QA).

To address long-range dependencies, prior work has proposed increasingly sophisticated memory architectures. Hierarchical frameworks [31, 32] pre-compute multi-scale temporal representations to enable structured retrieval, while graph-based methods [25, 33] explicitly model relational dependencies across events. These designs are motivated by the assumption that naive segment-level retrieval is insufficient for capturing high-level semantics (Fig. 1(a)).

However, constructing structured memories is not only computationally heavy but also reasoning-demanding. For instance, EgoRAG [31] builds a hierarchical memory by progressively merging 30-second segments into hour-level and day-level summaries, requiring multiple rounds of MLLM inference at each temporal scale. WorldMM [33] goes further: it constructs multi-scale knowledge graphs across different temporal resolutions, where each scale requires caption generation, triplet extraction, and graph consolidation via a Large Language Model (LLM) that continuously merges new knowledge while resolving conflicts with existing relations. As video length increases, these multi-stage processes grow rapidly in both computational cost and MLLM calls. More importantly, these structuring decisions are made without access to the downstream query, forcing a query-agnostic abstraction that may not align with what is ultimately asked.

This reveals a practical issue of *Intelligence Allocation*. If strong MLLMs are available, using them primarily for query-agnostic preprocessing is inefficient: it expends high-capability reasoning before the task is even specified. A more principled alternative is to preserve high-fidelity episodic evidence and defer semantic composition to inference time, when the query determines which relations and abstractions matter. From this perspective, the bottleneck shifts from building static semantics in advance to enabling high-recall retrieval that consistently delivers the right evidence.

Motivated by this view, we adopt a simpler design (Fig. 1(b)): rather than constructing complex semantic hierarchies during memory building, we retrieve over minimally processed episodic segments and form task-specific relations on demand. To this end, we introduce MERIT, a lightweight yet effective framework built on two key ideas: multi-key indexing and neighbor filtering.

First, we propose a **Multi-key Indexing** strategy. Rather than constructing hierarchical or graph-based structures, we attach multiple complementary keys to each minimal temporal segment. These keys capture different perspectives—such as event-centric, object-centric, dialogue-centric, or summary-level cues—allowing diverse queries to match relevant evidence without requiring explicit multi-scale memory design. By enriching each fine-grained segment with heterogeneous semantic anchors, we substantially improve retrieval robustness while keeping memory construction simple.

Second, we introduce **Neighbor Filtering**, a query-aware local aggregation mechanism grounded in the inductive bias of temporal continuity. When a segment is retrieved via key matching, its temporally adjacent neighbors are jointly considered to form a local evidence cluster. Within this cluster, we perform query-conditioned relevance selection to extract information most pertinent to the query. In other words, neighbor filtering consists of both local temporal expansion and query-aware evidence refinement, enabling coherent context reconstruction without pre-computed hierarchical structures.

Through these two simple components, our framework eliminates the need for expensive multi-stage memory consolidation while maintaining strong retrieval quality. As shown in Fig. 1(c), MERIT achieves state-of-the-art performance across multiple long-video QA benchmarks, including EgoLifeQA [31], LVBench [27] and Video-MME [8], outperforming prior methods such as EgoRAG [31] and WorldMM [33] with a significantly simpler memory pipeline. These results validate our central hypothesis: deferring semantic composition to query time, rather than investing in query-agnostic preprocessing, yields both better performance and greater efficiency.

2 Related Work

2.1 Ultra-long Video Understanding with MLLMs

The paradigm of video understanding has shifted from short-clip analysis to extended temporal reasoning. While recent proprietary MLLMs [6, 20] and open-source models [3, 5, 15, 24, 34, 37, 38] can natively process hour-long videos via extended context windows, the emergence of ultra-long benchmarks [31, 32] has pushed context requirements beyond these capacities. Despite increased window sizes, processing such extreme scales remains computationally prohibitive. Standard MLLMs under strict memory limits often resort to sparse uniform sampling, which inevitably discards fine-grained details and misses critical events.

Our Work. Rather than modifying or retraining the MLLM backbone, MERIT leverages off-the-shelf MLLMs by providing them with only query-relevant evidence. Since these models are highly capable yet constrained by limited context

windows, the key is to feed them the evidence that matters for each query rather than the entire video. To this end, MERIT avoids heavy query-agnostic memory construction and instead maintains a lightweight memory for inference time retrieval and evidence curation. This defers query-specific reasoning to inference, where the model's capacity can be used most effectively.

2.2 Memory-based Architectures for Video QA

To overcome the context window limitations of MLLMs, recent studies [11, 12, 16–18, 23, 25, 29–33] have adopted the Retrieval-Augmented Generation (RAG) paradigm for the video domain. These frameworks construct a structured external memory from video data, enabling the system to retrieve and incorporate relevant visual evidence at inference time.

Hierarchy-based Memory. Temporal hierarchy is a widely adopted structure for managing long-form video memory. EgoRAG [31] organizes video data into multiple temporal scales by recursively summarizing short 30-second segments. While this multi-level indexing provides a broad overview, the recursive summarization process inevitably discards fine-grained details. Furthermore, such top-down retrieval is highly sensitive to initial errors; a mismatch at the coarse summary level often leads to total QA failure. To mitigate these structural weaknesses, Ego-R1 [32] introduces an agentic approach that utilizes multi-turn tool calling to dynamically navigate the hierarchical memory. However, performance remains fundamentally limited by the resolution of the underlying summaries.

Graph-based Memory. Another prominent direction involves representing video memory as a graph to capture complex relational context. Techniques from text-based RAG [10, 11] have been extended to the multimodal domain [12], incorporating visual features into structured knowledge graphs. Models such as HippoRAG [11] and HippoMM [16] utilize short-to-long-term memory consolidation to build semantic graphs, while M3-Agent [17] incorporates entity-centric episodic and semantic memory. WorldMM [33] further utilizes multiple graph-based memories for adaptive retrieval, and EGAgent [22] constructs time-aware graphs to improve temporal reasoning. However, graph-based approaches require intensive pre-computation for graph construction, incurring significant computational overhead and latency. For instance, WorldMM [33] not only constructs multi-granular episodic graphs but also aggregates them into a global semantic graph, which inevitably necessitates continuous memory consolidation as the representation expands. Consequently, managing such cascading structural updates becomes prohibitively expensive as video length scales to multiple days.

Our Work. In contrast to pre-structured hierarchies or graphs, MERIT utilizes a multi-key episodic memory that bypasses the intensive preprocessing required for complex memory construction. By maintaining simple yet effective keys, we preserve fine-grained episodic details while minimizing memory building costs. Our framework achieves both efficiency and semantic depth by performing on-demand temporal expansion during inference, providing a scalable solution for ultra-long video understanding.

2.3 Caption-based Video RAG

An alternative paradigm represents video understanding in the language space, converting visual content into captions and delegating reasoning to a powerful LLM. LLoVi [36] densely captions short clips and aggregates them with an LLM, SiLVR [35] extends this caption-and-reason recipe with multisensory descriptions fed into a dedicated reasoning LLM, and VideoTree [28] organizes frames into a query-adaptive hierarchical tree for coarse-to-fine reasoning. As videos grow longer, video RAG systems build on such textual representations and retrieve only the query-relevant evidence from them at inference time [2, 18, 30]. Since retrieval is performed after the query is given, several works further generate query-aware, more informative evidence at this stage. For instance, iRAG [1] keeps only a lightweight index after preprocessing. For each query, it runs heavier vision models on the retrieved clips to caption details that were not extracted beforehand. DrVideo [19], instead, starts from a coarse document and runs an agentic loop, re-captioning key frames until the gathered information suffices.

Our Work. MERIT is likewise a caption-based, agentic video RAG framework. To obtain query-aware evidence, however, MERIT does not re-run captioning on the raw video at every inference step. Instead, it retrieves diverse relevant clips through complementary multi-keys enabling high-recall retrieval, and then distills query-aware, rich information from the pre-built captions via neighbor filtering. In doing so, MERIT also reconstructs the surrounding temporal context that single-clip retrieval would otherwise miss, without re-captioning per query.

3 Method

In this section, we present MERIT, a minimalist agentic framework designed to improve retrieval accuracy through a straightforward and simplified memory structure for ultra-long video QA. Unlike prior approaches [33] that rely on complex hierarchical graph construction, our method maintains minimal episodic representations and performs query-driven temporal expansion on demand.

3.1 Problem Formulation

In the ultra-long video setting, directly performing question answering (QA) by feeding raw video frames into MLLMs is infeasible due to context length limitations. Therefore, the common paradigm constructs an external memory representation M from the video V:

$$M = \mathcal{F}(V), \tag{1}$$

where $\mathcal{F}(\cdot)$ denotes the memory construction procedure. The memory is built in a query-agnostic manner, meaning that the natural language query Q is not available during memory construction. At inference time, the QA solver takes the query Q and the pre-built memory M to produce the final answer A:

$$A = \mathcal{G}(Q, M), \tag{2}$$

where $\mathcal{G}(\cdot)$ denotes the QA solver.

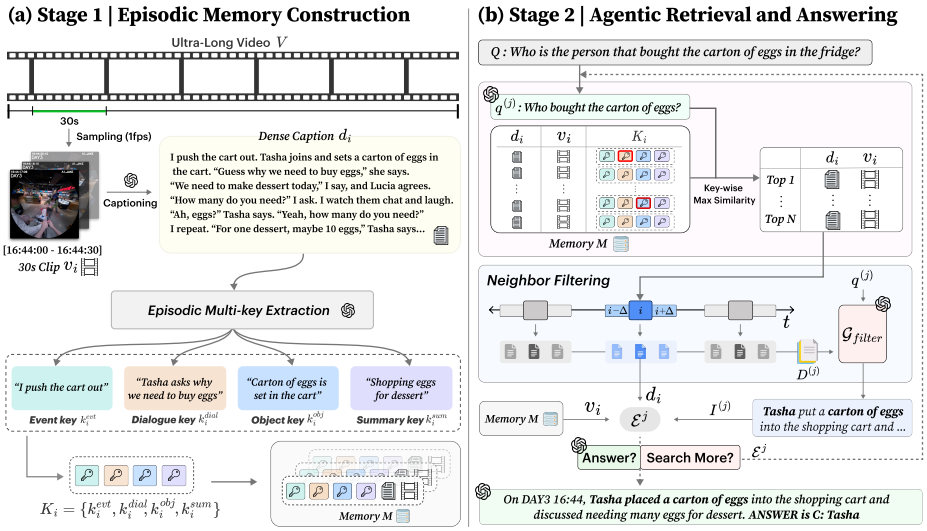

Fig. 2: Overall Pipeline of MERIT. (a) **Stage 1: Episodic memory construction.** For each 30-second video clip, dense captions are generated. Subsequently, an episodic multi-key extraction process derives four distinct keys per clip, collectively forming the memory M. (b) **Stage 2: Agentic retrieval and answering.** The solver regenerates a query and matches it against multi-keys to retrieve the most relevant clips. Neighbor Filtering then expands the temporal context around these clips, extracting additional query-relevant information to formulate the final answer.

3.2 Overall Pipeline

Fig. 2 illustrates the overall pipeline of our method MERIT. MERIT follows a minimalist memory-based agentic pipeline, with an emphasis on high-recall retrieval and lightweight preprocessing.

Episodic memory construction. We partition the video V into a sequence of non-overlapping clips: $V = \{v_1, v_2, \ldots, v_T\}$. We generate a dense caption d_i for each clip v_i. For each clip, we additionally derive a set of textual keys K_i. The resulting memory is:

$$M = \mathcal{F}(V) = \{(v_i, d_i, K_i)\}_{i=1}^{T}. \tag{3}$$

Conceptually, M can be viewed as a minimal key–value store, where K_i serves as retrieval keys and (v_i, d_i) provides the associated values.

Agentic retrieval and answering. During QA, the solver $\mathcal{G}$ runs an iterative multi-turn procedure [17, 32, 33]. At round j, the solver forms a retrieval query $q^{(j)}$ from Q and the retrieval results accumulated up to the previous round, namely *evidence* $\mathcal{E}^{(j-1)}$, with $\mathcal{E}^{(0)} = \emptyset$. It then retrieves a set of relevant clip indices $R^{(j)} \subseteq \{1, \ldots, T\}$ by matching $q^{(j)}$ against the stored keys $\{K_i\}$ using a sentence embedding model $E(\cdot)$ and similarity scoring. For each retrieved index $i \in R^{(j)}$, the agent collects the corresponding episodic record as an evidence item

$e_i = (v_i, d_i)$, where d_i is used during the iterative retrieval, and v_i is reserved for generating the final answer A. It updates the accumulated evidence set as

$$\mathcal{E}^{(j)} = \mathcal{E}^{(j-1)} \cup \{e_i \mid i \in R^{(j)}\}. \tag{4}$$

After each retrieval round, the agent assesses whether the currently collected evidence is sufficient to answer Q. If not, it refines the retrieval query and repeats retrieval. Once sufficient evidence is obtained, the agent leverages the solver's reasoning to integrate evidence across retrieved episodes and infer the relationships required to produce A.

3.3 Multi-Key Memory and Retrieval

MERIT defers query-specific relation reasoning to the solver at inference time, so retrieval becomes the key bottleneck, as the solver can only reason over relationships present in the retrieved evidence $\mathcal{E}$. In long video QA, a clip can be relevant via different cues, such as actions, spoken mentions, object interactions and state changes, or coarse scene context; collapsing these cues into a single textual index is often brittle. Therefore, instead of the standard single-key indexing, we represent each record (v_i, d_i) with a *set* of complementary keys K_i and use late-interaction matching [13, 21, 26] to favor high-recall retrieval.

Multi-Key Memory. The representation is explicitly formulated as a combination of four distinct categories, defined as

$$K_i = \{k_i^{evt}, k_i^{dial}, k_i^{obj}, k_i^{sum}\}. \tag{5}$$

- Event / Action Key (k_i^{evt}): Captures observable physical actions and interactions between entities.
- Dialogue / Mention Key (k_i^{dial}): Records spoken content, exact words, or commands mentioned in the audio track, providing crucial linguistic context.
- Object Key (k_i^{obj}): Describes specific items being handled, requested, or moved, along with their state transitions.
- Summary Key (k_i^{sum}): A compact keyword-style abstraction capturing the core narrative and the coarse information of the overall clip.

This decoupled representation ensures that each clip can be matched from diverse perspectives, effectively handling varying query intents.

Maximum Similarity Retrieval. At retrieval round j, given the retrieval query $q^{(j)}$, we embed $q^{(j)}$ and each key $k \in K_i$ using $E(\cdot)$ and define the clip relevance by the maximum cosine similarity:

$$S_i^{(j)} = \max_{k \in K_i} \mathrm{sim}\big(E(q^{(j)}),\, E(k)\big), \tag{6}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes the cosine similarity score between two embeddings. We then select the top N clips:

$$R^{(j)} = \mathrm{TopN}\big(\{S_i^{(j)}\}_{i=1}^{T}\big). \tag{7}$$

This maximum similarity formulation improves recall by allowing each clip to match the query through its best-aligned aspect.

3.4 Temporal Expansion via Neighbor Filtering

Videos exhibit strong temporal locality: the evidence required to answer a query often spans multiple nearby moments rather than a single 30-second clip. While short clips enable lightweight indexing, retrieving only the anchor clip can miss necessary preconditions or follow-up context, limiting the solver's ability to infer query-specific relationships.

Neighbor Filtering. We exploit this inductive bias with a minimal temporal structure. Instead of pre-computing multi-granularity hierarchies, we attach a local neighborhood of temporally adjacent clips to each retrieved anchor on demand, effectively forming the simplest graph over the timeline. This query-driven expansion enriches the retrieved evidence at low cost, allowing a powerful solver to compose higher-level temporal and semantic relations in a query-aware manner at inference time.

For each retrieved index $i \in R^{(j)}$ in round j, we define a symmetric temporal window

$$W_i = \{t \in \{1, ..., T\} | i - \Delta \leq t \leq i + \Delta\}, \tag{8}$$

where Δ is an integer radius indicating the number of adjacent clips on each side (e.g., Δ=2 spans ±1 minute under a 30-second clip setting).

The full textual context for neighbor filtering at round j is obtained by concatenating the dense captions within the neighborhood of each retrieved index:

$$D^{(j)} = \text{Concat}\left(\{d_t | t \in W_i, i \in R^{(j)}\}\right). \tag{9}$$

We then use the solver $\mathcal{G}$ as a query-aware filter to distill only the information relevant to Q from this expanded context: $I^{(j)} = \mathcal{G}_{\text{filter}}(Q, D^{(j)})$, where $\mathcal{G}_{\text{filter}}$ denotes the solver equipped with a filtering prompt. The resulting $I^{(j)}$ is appended to the evidence set $\mathcal{E}^{(j)}$ and used for subsequent query-aware reasoning.

4 Experiments

4.1 Benchmarks

We evaluate MERIT against existing memory-based models on three long-video benchmarks, formulating all tasks as multiple-choice questions (MCQs) and using QA accuracy as the primary evaluation metric.

EgoLifeQA [31]: An ultra-long egocentric dataset averaging 44.3 hours per video, capturing six individuals over seven days with 500 QA pairs. It features five QA types: EntityLog and EventRecall test precise factual grounding of objects and events, while HabitInsight, RelationMap, and TaskMaster evaluate semantic reasoning over human behaviors, interactions, and future planning.

LVBench [27]: A benchmark for hour-long video understanding with an average duration of 1.12 hours and a maximum of 2.33 hours. It includes 1,549 questions across six skill categories: Entity Recognition (ER) and Event Understanding (EU) for tracking and classification; Temporal Grounding (TG) and Key Information Retrieval (KIR) for locating specific moments and details; and

Table 1: Comparison of MERIT against various baselines on EgoLifeQA [31]. Baseline results are from [22, 33]. **Bold** and underline denote the best and second-best scores, respectively. Solvers are indicated per model (pre-trained models report weights only).

Model	EntityLog	EventRecall	HabitInsight	RelationMap	TaskMaster	Avg
MLLMs (Uniform Sampling)						
Qwen3-VL-8B [3]	35.2	30.2	39.3	46.4	46.0	38.6
Gemini 2.5 Pro [6]	43.2	40.5	41.0	55.2	52.4	46.4
GPT-5 [20]	47.2	42.1	47.5	53.6	55.6	48.6
Hierarchical Memory Based						
EgoRAG [31] (GPT-5)	40.0	56.3	62.3	54.4	52.4	52.0
Ego-R1 [32] (3B)	51.2	53.2	63.9	50.4	50.8	53.0
Graph Memory Based						
LightRAG [10] (GPT-5)	40.8	48.4	67.2	50.4	44.4	48.8
HippoRAG [11] (GPT-5)	48.8	60.3	70.5	60.8	66.7	59.6
Video-RAG [18] (GPT-5)	49.6	56.3	67.2	55.2	54.0	55.4
HippoMM [16] (GPT-5)	45.6	53.2	70.5	55.2	58.7	54.6
M3-Agent [17] (7B)	44.4	54.8	62.3	56.8	54.0	53.5
EGAgent [22] (Gemini 2.5 Pro)	54.4	57.1	60.3	62.4	**74.6**	57.5
WorldMM [33] (Qwen3-VL-8B)	49.6	56.4	63.9	58.4	58.7	56.4
WorldMM [33] (GPT-5)	62.4	64.3	**75.4**	62.4	71.4	65.6
Ours						
MERIT (Qwen3-VL-8B)	43.2	54.0	67.2	53.6	65.1	54.2
MERIT (Gemini 2.5 Pro)	60.8	61.1	65.6	65.6	73.0	64.2
MERIT (GPT-5)	**67.2**	**70.6**	73.8	**74.4**	71.4	**71.2**

Summarization (Sum) and Reasoning (Rea) for synthesizing global content and causal inference. Among these, ER and EU account for the largest proportion. **Video-MME [8]**: This is a comprehensive benchmark for evaluating the general video understanding capabilities of MLLMs. It is categorized into Short, Medium, and Long subsets based on video duration. In our experiments, we use only the Long subset (30–60 mins), which comprises 900 questions from 300 videos, with three questions per video. The benchmark covers 12 question types, with Object Reasoning, Action Reasoning, and Information Synopsis being the most prevalent in the Long subset.

4.2 Implementation Details

To evaluate our framework, we applied MERIT in Qwen3-VL-8B [3], GPT-5 [20], and Gemini 2.5 Pro [6]. Following the previous work [33], we employ GPT-5-mini [20] to generate dense captions for each 30-second clip. We also use the same model to generate episodic multi-key for each segment. For EgoLifeQA [31] and Video-MME(Long) [8], we incorporate ASR transcripts as an additional modality during captioning [22, 33]. For LVBench [27], which does not rely on dialogue or speech, we use only visual frames for captioning and exclude the dialogue key. The prompts used for multi-key extraction and neighbor filtering, along with further implementation details, are provided in the Appendix.

Table 2: Comparison of MERIT against various baselines on LVBench [27] and Video-MME(L) [8] benchmarks. Baseline results are from [22, 33]. Solvers are indicated per model (pre-trained models report weights only).

Model	LVBench	Video-MME(L)
MLLMs (Uniform Sampling)		
Gemini 2.5 Pro [6]	57.0	55.7
GPT-5 [20]	60.4	74.3
Hierarchical Memory Based		
EgoRAG [31] (GPT-5)	32.2	41.1
Ego-R1 [32] (3B)	34.1	42.7
Graph Memory Based		
LightRAG [10] (GPT-5)	30.4	46.6
HippoRAG [11] (GPT-5)	54.0	52.1
Video-RAG [18] (GPT-5)	33.1	55.4
HippoMM [16] (GPT-5)	38.2	41.6
M3-Agent [17] (7B)	49.3	55.3
EGAgent [22] (Gemini 2.5 Pro)	-	74.1
WorldMM [33] (GPT-5)	61.9	76.6
Ours		
MERIT (Gemini 2.5 Pro)	-	76.3
MERIT (GPT-5)	**71.8**	**77.7**

4.3 Main Results

Results on Ultra-Long Video QA. Table 1 presents the evaluation on the ultra-long benchmark EgoLifeQA. Uniform-sampling MLLMs exhibit the lowest overall performance due to strict context length limits. This confirms that processing full ultra-long videos is computationally prohibitive and inaccurate, thereby necessitating memory-based architectures. MERIT outperforms existing memory-based approaches and establishes a new state-of-the-art. Using GPT-5, our framework achieves an average accuracy of 71.2%, yielding a +5.6% improvement over the previous SOTA method.

An analysis of individual QA types highlights the strengths of MERIT in fine-grained categories such as EntityLog and EventRecall. This proves that our simple key representation prevents information loss during retrieval, passing intact raw evidence directly to the solver. Furthermore, the significant +12.0% improvement in RelationMap proves that expanding the temporal radius during inference successfully captures query-relevant semantic information without the heavy overhead of pre-structuring semantic memory. This superiority is consistent across advanced models; utilizing Gemini 2.5 Pro yields a +6.7% gain over the best graph-memory baseline. Consequently, these results indicate that MERIT effectively leverages advanced reasoning to maximize performance.

Results on Hour-Long Video QA. Table 2 demonstrates that MERIT also establishes new SOTA results on hour-long benchmarks. Specifically, on LVBench,

Table 3: Ablation study on Neighbor Filtering (NF) across two solver models. *Target Time Oracle* provides ground-truth temporal segments as an upper bound, while MERIT uses retrieved segments.

EXP	NF	EntityLog	EventRecall	HabitInsight	RelationMap	TaskMaster	Avg	Hit Rate
Target Time Oracle (Qwen3-VL-8B [3])	✗	**60.0**	**69.8**	**70.5**	54.4	73.0	64.0	1.0
	✓	56.8	66.7	68.9	**67.2**	**76.2**	**65.8**	1.0
MERIT (Qwen3-VL-8B [3])	✗	41.6	**54.0**	50.8	42.4	57.1	48.0	0.35
	✓	**43.2**	**54.0**	**67.2**	**53.6**	**65.1**	**54.2**	**0.42**
Target Time Oracle (GPT-5 [20])	✗	**82.4**	**89.7**	**85.3**	73.6	82.5	82.4	1.0
	✓	80.8	88.1	**85.3**	**76.8**	**88.9**	**83.2**	1.0
MERIT (GPT-5 [20])	✗	**68.0**	69.8	68.9	73.6	68.3	70.0	0.40
	✓	67.2	**70.6**	**73.8**	**74.4**	**71.4**	**71.2**	**0.56**

our GPT-5 setting achieves 71.8%, outperforming the prior SOTA by a significant margin of +9.9%. Unlike the day-long setting, uniform-sampling MLLMs show competitive performance on these shorter videos (averaging 0.69 hours for Video-MME(L) and 1.12 hours for LVBench), occasionally surpassing existing memory-based models. However, as video duration increases to over an hour in LVBench, the performance gap between uniform sampling and MERIT widens significantly. For instance, with GPT-5, the gap expands from +3.4% on Video-MME(L) to an impressive +11.4% on LVBench. This confirms that while MLLMs can manage moderately long contexts, their scalability strictly limits them in longer settings, highlighting the robustness and the efficacy of our memory framework for arbitrarily long videos.

4.4 Ablation Studies

We conduct ablation studies to evaluate the individual components of MERIT. All ablation experiments are conducted on the EgoLifeQA [31] benchmark. For the backbone QA solvers, we use both open-source and proprietary models. To analyze retrieval performance, we additionally measure *Hit Rate*. Given a query and its corresponding target timestamp, a hit is scored as 1 if the target time falls within the retrieved memory interval, and 0 otherwise. The Hit Rate is the average of hit scores across all queries.

Effect of Neighbor Filtering. Table 3 evaluates the impact of expanded temporal context on QA performance. Neighbor Filtering (NF) is a mechanism that dynamically expands the temporal radius around a retrieved clip at inference time to incorporate surrounding information. In the absence of NF, the solver is strictly limited to the memory of the single retrieved 30-second clip, without any adjacent context.

To establish an upper bound for this single-clip setting, we evaluate a "Target Time Oracle" configuration, which directly provides the solver with frames and captions from the ground-truth 30-second clip. This oracle setting reveals a substantial performance gap compared to prior state-of-the-art models, confirming that precise retrieval remains the primary bottleneck in ultra-long video QA.

Table 4: Performance analysis based on key combinations with Qwen3-VL-8B [3]. Darker colors indicate higher accuracy.

N Key	Event	Dial	Object	Sum	Acc	Avg
1 key	✓				44.6	46.4
		✓			44.8	
			✓		46.8	
				✓	49.4	
2 key	✓	✓			49.6	49.4
	✓		✓		47.0	
	✓			✓	49.0	
		✓	✓		49.2	
		✓		✓	51.2	
			✓	✓	50.6	
3 key	✓	✓	✓		49.8	51.1
	✓	✓		✓	52.0	
	✓		✓	✓	50.8	
		✓	✓	✓	51.6	
4 key	✓	✓	✓	✓	**54.2**	54.2

Table 5: Retrieval accuracy comparison with baseline models using GPT-5 [20]. The (*) denotes values obtained from our reproduction of the baseline models under the same evaluation settings.

EXP	QA Acc	Hit Rate
EgoRAG*	49.7	0.15
WorldMM*	63.2	0.53
MERIT	**71.2**	**0.56**

Table 6: Comparison of memory construction efficiency with reproduced baseline models(*) using GPT-5 [20].

EXP	LLM Call Count	Input Tokens	Output Tokens
EgoRAG*	**339**	1,049K	**247K**
WorldMM*	23,731	6,860K	2,472K
MERIT	6,223	**815K**	288K

Applying NF mitigates this bottleneck by expanding the temporal context during inference, which significantly improves both the hit rate and overall accuracy. Specifically, the expanded temporal scope drives consistent performance gains in QA categories that inherently require a broader semantic context, such as HabitInsight, RelationMap, and TaskMaster.

Notably, these improvements are also consistently observed in the "Target Time Oracle w/ NF" setting. This indicates that even with perfect temporal localization, a single isolated 30-second segment often lacks sufficient episodic context to resolve complex queries. By incorporating information from surrounding clips, NF provides the necessary temporal and relational cues to capture a complete semantic understanding. Consequently, this on-demand expansion yields significant performance gains without incurring the massive computational overhead of additional pre-processing.

Effect of Multi-Key Combinations. Table 4 details the performance variations across different configurations of our multi-key representation using the Qwen3-VL-8B [3]. Specifically, our framework extracts four distinct keys per clip: event, dialogue, object, and summary. This ablation study evaluates the impact of these individual keys, their various combinations, and the effect of scaling the total number of keys utilized for retrieval.

Under the single-key setting, the summary key outperforms the other individual keys by a margin of up to +4.8%. Instead of focusing on specific aspects, the summary key captures coarse, abstract context, serving as a robust primary anchor for retrieval.

Furthermore, combining multiple keys leads to a steady increase in overall accuracy. While intermediate multi-key combinations show minor variances, the distinct keys function in a highly complementary manner. Consequently, utilizing all four keys simultaneously yields the peak accuracy of 54.2%. These findings empirically validate our premise: maintaining a decoupled multi-key representation ensures that each video clip can be accurately matched from diverse semantic perspectives, thereby effectively handling varying query intents.

Retrieval Accuracy vs. Baselines. To investigate the correlation between retrieval accuracy and overall QA performance, we evaluate the hit rate of reproduced models representing distinct memory architectures (Table 5). These baselines are built on the same 30-second dense captions as MERIT, produced by the same captioner [20], using their official code for controlled comparison.

Compared to the hierarchical memory of EgoRAG [31] and the graph memory of WorldMM [33], MERIT employs the simplest memory key structure yet achieves both the highest hit rate and the highest overall accuracy. This direct alignment confirms that successful information retrieval is critical for QA performance. Unlike MERIT and EgoRAG, which retrieve in a fixed 30-second granularity, WorldMM retrieves clips across multi-scale temporal windows, including up to 1 hour segments. While retrieving such extensive temporal windows inherently increases the probability of capturing the target timestamp, our framework still achieves a higher hit rate using only a single granularity. In our reproduction, evaluating WorldMM solely with 30-second clips yields a hit rate of 0.31. These results demonstrate that precise, fine-grained retrieval capabilities are closely correlated with maximizing overall QA performance.

Memory Construction Efficiency. In Table 6, we evaluate the memory construction cost. Since all methods use the same 30-second captions, we compare memory construction efficiency by measuring the total number of caption tokens consumed by each method, excluding the common initial captioning step.

Compared to WorldMM [33], MERIT demonstrates a significant efficiency advantage, achieving an **8.4**× reduction in input tokens and an **8.6**× reduction in output tokens. WorldMM's pipeline generates captions across four granularities, applies OpenIE to each, and continuously updates multi-level semantic graphs, incurring 23.7K LLM calls and millions of input/output tokens. In contrast, MERIT builds its memory in a single pass: each clip is captioned once and indexed with lightweight multi-keys, which eliminates the repeated generation and graph-update steps.

Furthermore, while EgoRAG [31] requires fewer LLM calls by batching inputs within its hierarchical structure, MERIT consumes 22% fewer input tokens (815K vs. 1.05M). This efficiency is achieved because MERIT processes each caption exactly once, avoiding the redundant re-reading of intermediate summaries inherent to EgoRAG's hierarchy levels. Although MERIT generates slightly more output tokens than EgoRAG, this is a deliberate design choice: rather than heavily compressing information into single summaries, MERIT generates fine-grained multi-keys per clip, which is critical for preserving detailed temporal context and enabling highly accurate retrieval.

Table 7: Impact of the retrieved clip count (Top N) across both open-source and proprietary QA solvers. Colors denote higher performance within each individual solver.

Solver	Top N	EntityLog	EventRecall	HabitInsight	RelationMap	TaskMaster	Avg
Qwen3-VL-4B [3]	5	41.6	47.6	63.9	37.6	55.6	**46.6**
	10	37.6	53.2	59.0	39.2	52.4	46.4
Qwen3-VL-8B [3]	5	43.2	54.0	67.2	53.6	65.1	**54.2**
	10	44.0	57.9	63.9	40.8	68.3	52.2
Gemini 2.5 pro [6]	5	53.6	62.7	60.7	65.6	66.7	61.4
	10	60.8	61.1	65.6	65.6	73.0	**64.2**
GPT-5 [20]	5	62.4	67.5	73.8	68.0	69.8	67.4
	10	67.2	70.6	73.8	74.4	71.4	**71.2**

Effect of Retrieved Clip Quantity Across Solver Capacities. Table 7 investigates the effect of the number of retrieved clips (Top N) across QA solvers with varying capacities. We observe a distinct divergence in performance trends dependent on the solver's inherent capabilities. For open-source models [3], restricting retrieval to Top 5 clips yields optimal results, whereas expanding to Top 10 degrades performance. This suggests that models with limited reasoning capacity struggle to effectively consolidate information from extended contexts, often becoming confused by the increased noise during inference.

Conversely, proprietary models [6, 20] demonstrate substantial improvements when provided with a larger pool of clips. Increasing to Top 10 boosts the accuracy of Gemini 2.5 Pro (61.4% to 64.2%) and GPT-5 (67.4% to 71.2%). This suggests that high-capacity solvers possess the noise-tolerance required to effectively filter query-aware evidence from expanded context. Rather than being hindered by redundant information in additional temporal windows, they distill critical semantic cues from extended contexts to enhance reasoning accuracy.

5 Qualitative Results

Effect of Multi-Key Memory. Fig. 3 (a) illustrates an example from the EntityLog category of EgoLifeQA [31]. The query requires identifying the specific moment an object was handed over. Our multi-key memory matches this query to the stored event key: *"I hand Tasha a bottle of AD calcium milk."* By retrieving the exact target segment associated with this key, our retrieval mechanism provides the solver with the visual evidence necessary to identify the target entity. This successful retrieval of the fine-grained event leads to the correct answer.

Effect of Neighbor Filtering. Fig. 3 (b) presents a RelationMap case from EgoLifeQA [31], requiring an understanding of interpersonal relationships and interactions. The query asks about an event from two days before the query time, where multiple individuals are present. While multi-key matching accurately retrieves the target segment, a single clip lacks the context to identify specific individuals. Neighbor filtering resolves this by expanding the temporal context,

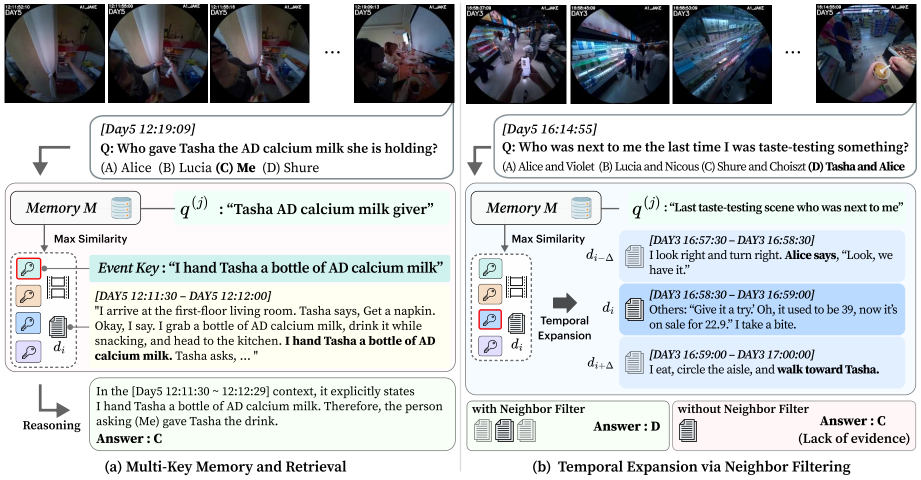

Fig. 3: Qualitative results of MERIT on EgoLifeQA [31]. (a) The event key precisely matches the query, enabling retrieval of the exact target segment. (b) While both settings retrieve the correct segment, neighbor filter supplies the surrounding $\pm\Delta$ context necessary to resolve relational queries that cannot be answered from a single clip alone.

allowing the query-aware filter to extract the details: *"I was taste-testing and next to me were Tasha and Alice."* This confirms that while multi-key indexing localizes anchor clips, neighbor filtering provides the broader situational context required for relational queries—without relying on pre-built semantic memories.

6 Conclusion

We present MERIT, a simple yet effective framework for ultra-long video understanding that preserves fine-grained episodic evidence through multi-key indexing and enriches retrieved context via query-time neighbor filtering. Despite its simplicity, MERIT achieves state-of-the-art performance across multiple benchmarks, suggesting that deferring semantic composition to inference time is more effective than query-agnostic preprocessing. We hope MERIT serves as a practical and scalable foundation for future work on ultra-long video understanding.

Acknowledgements

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2020-II201361, Artificial Intelligence Graduate School Program (Yonsei University)), National Research Foundation of Korea (NRF) (RS-2025-00554790), and No.RS-2022-II220124, Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities.

References

1. Arefeen, M.A., Debnath, B., Uddin, M.Y.S., Chakradhar, S.: irag: Advancing rag for videos with an incremental approach. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. pp. 4341–4348 (2024)
2. Ataallah, K., Shen, X., Abdelrahman, E., Sleiman, E., Zhuge, M., Ding, J., Zhu, D., Schmidhuber, J., Elhoseiny, M.: Goldfish: Vision-language understanding of arbitrarily long videos. In: European Conference on Computer Vision. pp. 251–267. Springer (2024)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
6. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
7. Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R.O., Larson, J.: From local to global: A graph rag approach to query-focused summarization. arXiv preprint arXiv:2404.16130 (2024)
8. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24108–24118 (2025)
9. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., Wang, H.: Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 (2023)
10. Guo, Z., Xia, L., Yu, Y., Ao, T., Huang, C.: Lightrag: Simple and fast retrieval-augmented generation. arXiv preprint arXiv:2410.05779 **2**(3) (2024)
11. Gutiérrez, B.J., Shu, Y., Qi, W., Zhou, S., Su, Y.: From rag to memory: Non-parametric continual learning for large language models. arXiv preprint arXiv:2502.14802 (2025)
12. Jeong, S., Kim, K., Baek, J., Hwang, S.J.: Videorag: Retrieval-augmented generation over video corpus. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 21278–21298 (2025)
13. Khattab, O., Zaharia, M.: Colbert: Efficient and effective passage search via contextualized late interaction over bert. In: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval. pp. 39–48 (2020)
14. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems **33**, 9459–9474 (2020)

15. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
16. Lin, Y., Wang, Q., Ye, H., Fu, Y., Li, H., Chen, Y., et al.: Hippomm: Hippocampal-inspired multimodal memory for long audiovisual event understanding. arXiv preprint arXiv:2504.10739 (2025)
17. Long, L., He, Y., Ye, W., Pan, Y., Lin, Y., Li, H., Zhao, J., Li, W.: Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory (2025)
18. Luo, Y., Zheng, X., Li, G., Yin, S., Lin, H., Fu, C., Huang, J., Ji, J., Chao, F., Luo, J., et al.: Video-rag: Visually-aligned retrieval-augmented long video comprehension (2024)
19. Ma, Z., Gou, C., Shi, H., Sun, B., Li, S., Rezatofighi, H., Cai, J.: Drvideo: Document retrieval based long video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18936–18946 (2025)
20. OpenAI: Gpt-5 system card (August 2025), `https://openai.com/index/gpt-5-system-card/`
21. Reddy, A., Martin, A., Yang, E., Yates, A., Sanders, K., Murray, K., Kriz, R., de Melo, C.M., Van Durme, B., Chellappa, R.: Video-colbert: Contextualized late interaction for text-to-video retrieval. In: CVPR. pp. 19691–19701 (2025)
22. Rege, A., Sadhu, A., Li, Y., Li, K., Vinayak, R.K., Chai, Y., Lee, Y.J., Kim, H.J.: Agentic very long video understanding (2026)
23. Ren, X., Xu, L., Xia, L., Wang, S., Yin, D., Huang, C.: Videorag: Retrieval-augmented generation with extreme long-context videos. arXiv preprint arXiv:2502.01549 (2025)
24. Shen, X., Xiong, Y., Zhao, C., Wu, L., Chen, J., Zhu, C., Liu, Z., Xiao, F., Varadarajan, B., Bordes, F., Liu, Z., Xu, H., Kim, H.J., Soran, B., Krishnamoorthi, R., Elhoseiny, M., Chandra, V.: Longvu: Spatiotemporal adaptive compression for long video-language understanding. arXiv preprint arXiv:2410.17434 (2024)
25. Shen, X., Zhang, W., Chen, J., Elhoseiny, M.: Vgent: Graph-based retrieval-reasoning-augmented generation for long video understanding (2025)
26. Wan, D., Wang, H., Stengel-Eskin, E., Cho, J., Bansal, M.: Clamr: Contextualized late-interaction for multimodal content retrieval. arXiv preprint arXiv:2506.06144 (2025)
27. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Ding, M., Gu, X., Huang, S., Xu, B., et al.: Lvbench: An extreme long video understanding benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22958–22967 (2025)
28. Wang, Z., Yu, S., Stengel-Eskin, E., Yoon, J., Cheng, F., Bertasius, G., Bansal, M.: Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3272–3283 (2025)
29. Xu, Z., Zhang, J., Wang, Q., Liu, Y.: E-vrag: Enhancing long video understanding with resource-efficient retrieval augmented generation (2025)
30. Xue, Z., Zhang, J., Xie, X., Cai, Y., Liu, Y., Li, X., Tao, D.: Adavideorag: Omni-contextual adaptive retrieval-augmented efficient long video understanding. arXiv preprint arXiv:2506.13589 (2025)
31. Yang, J., Liu, S., Guo, H., Dong, Y., Zhang, X., Zhang, S., Wang, P., Zhou, Z., Xie, B., Wang, Z., et al.: Egolife: Towards egocentric life assistant. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28885–28900 (2025)
32. Yang, J., et al.: Ego-r1. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2025)

33. Yeo, W., Kim, K., Yoon, J., Hwang, S.J.: Worldmm: Dynamic multimodal memory agent for long video reasoning. arXiv preprint arXiv:2512.02425 (2025)
34. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., et al.: Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106 (2025)
35. Zhang, C., Lin, Y.B., Wang, Z., Bansal, M., Bertasius, G.: Silvr: A simple language-based video reasoning framework. arXiv preprint arXiv:2505.24869 (2025)
36. Zhang, C., Lu, T., Islam, M.M., Wang, Z., Yu, S., Bansal, M., Bertasius, G.: A simple llm framework for long-range video question-answering. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 21715–21737 (2024)
37. Zhang, Y., Li, B., Liu, h., Lee, Y.j., Gui, L., Fu, D., Feng, J., Liu, Z., Li, C.: Llava-next: A strong zero-shot video understanding model (April 2024), `https://llava-vl.github.io/blog/2024-04-30-llava-next-video/`
38. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Llava-video: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)