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] **Seon Joo Kim** [1]

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) (OpenAI, 2025; Comanici et al., 2025; Bai et al., 2025a;b; Zhang et al., 2024) has shifted video understanding from short clips to ultra-long streams spanning hours or 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, end-to-end modeling becomes infeasible, making external memory and Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) frameworks essential.

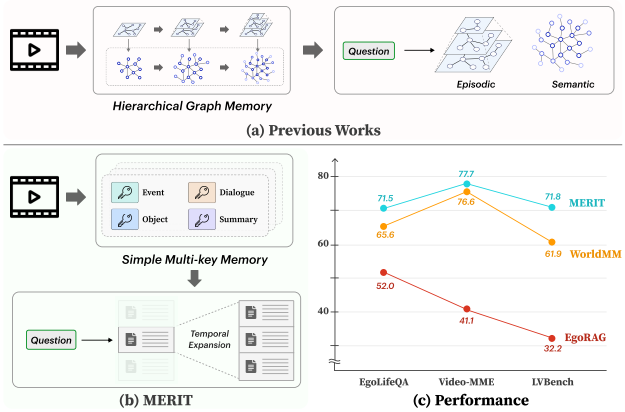


Figure 1. (a) Previous methods: Construct hierarchical graph-based episodic and semantic memories, and at inference time, searches and aggregate relevant information across these structured memories. (b) Our method: MERIT constructs a simple episodic multi-key memory, and at inference time, retrieves the matched clips along with temporal expansion to perform query-driven semantic reasoning. (c) Performance of MERIT against EgoRAG (Yang et al., 2025a) and WorldMM (Yeo et al., 2025) across various benchmarks.

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 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 (Yang et al., 2025a;b) pre-compute

[1]Yonsei University [2]Adobe Research. Correspondence to: Hyolim Kang <hyolimkang@yonsei.ac.kr>, Seon Joo Kim <seonjookim@yonsei.ac.kr>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

multi-scale temporal representations to enable structured retrieval, while graph-based methods (Shen et al., 2025; Yeo et al., 2025) 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 (Yang et al., 2025a) builds a hierarchical memory by progressively merging 30 seconds segments into hour-level and day-level summaries, requiring multiple rounds of MLLMs inference at each temporal scale. WorldMM (Yeo et al., 2025) goes further: it constructs multi-scale knowledge graphs across different temporal resolutions, where each scale requires caption generation, triplet extraction, and graph consolidation via an 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 MLLMs 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 will ultimately be 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 specified. A 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 temporal 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 question. In other words, neighbor filtering consists of both local temporal expansion and query-aware evidence refinement, enabling coherent context reconstruction without pre-computed hierarchies.

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(Yang et al., 2025a), LVBench (Wang et al., 2025) and Video-MME(Fu et al., 2025), outperforming prior methods such as EgoRAG(Yang et al., 2025a) and WorldMM(Yeo et al., 2025) 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

To overcome the context window limitations of MLLMs, recent studies (Luo et al., 2024; Xu et al., 2025; Ren et al., 2025; Xue et al., 2025; Shen et al., 2025; Long et al., 2025) 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.

Hierarchical Memory based Memory. Temporal hierarchy is a widely adopted structure for managing long-form video memory. EgoRAG (Yang et al., 2025a) organizes video data into multiple temporal scales by recursively summarizing short 30 seconds segments. While this multi-level indexing provides a broad overview, the 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 (Yang et al., 2025b) introduces an agentic approach that utilizes multi-turn tool calling to dynamically navigate the hierarchical memory. However, performance remains limited by the resolution of the underlying summaries.

Graph Memory based Memory. Another prominent direction involves representing video memory as a graph to capture complex relational context. Techniques from text-based RAG (Guo et al., 2024; Gutiérrez et al., 2025) have been extended to the multimodal domain (Jeong et al., 2025), incorporating visual features into structured knowl-

edge graphs. Models such as HippoRAG (Gutiérrez et al., 2025) and HippoMM (Lin et al., 2025) utilize short-to-long-term memory consolidation to build semantic graphs, while M3-Agent (Long et al., 2025) incorporates entity-centric episodic and semantic memory. WorldMM (Yeo et al., 2025) further utilizes multiple graph-based memories for adaptive retrieval, and EGAgent (Rege et al., 2026) 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 (Yeo et al., 2025) not only construct multi-granular episodic graphs but also aggregate 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.

3. Method

In this section, we present MERIT, a minimalist agentic framework to improve retrieval accuracy through a straightforward and simplified memory structure for ultra-long video QA. Unlike prior approaches (Yeo et al., 2025) that rely on complex hierarchical graph construction, our method maintains minimally processed 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 $\mathcal{G}(\cdot)$ takes the query Q and the pre-built memory M to produce the final answer A:

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

3.2. Overall Pipeline

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

3.2.1. 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.

3.2.2. Agentic retrieval and answering

During QA, the solver $\mathcal{G}$ runs an iterative multi-turn procedure. 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 the 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 to favor high recall retrieval.

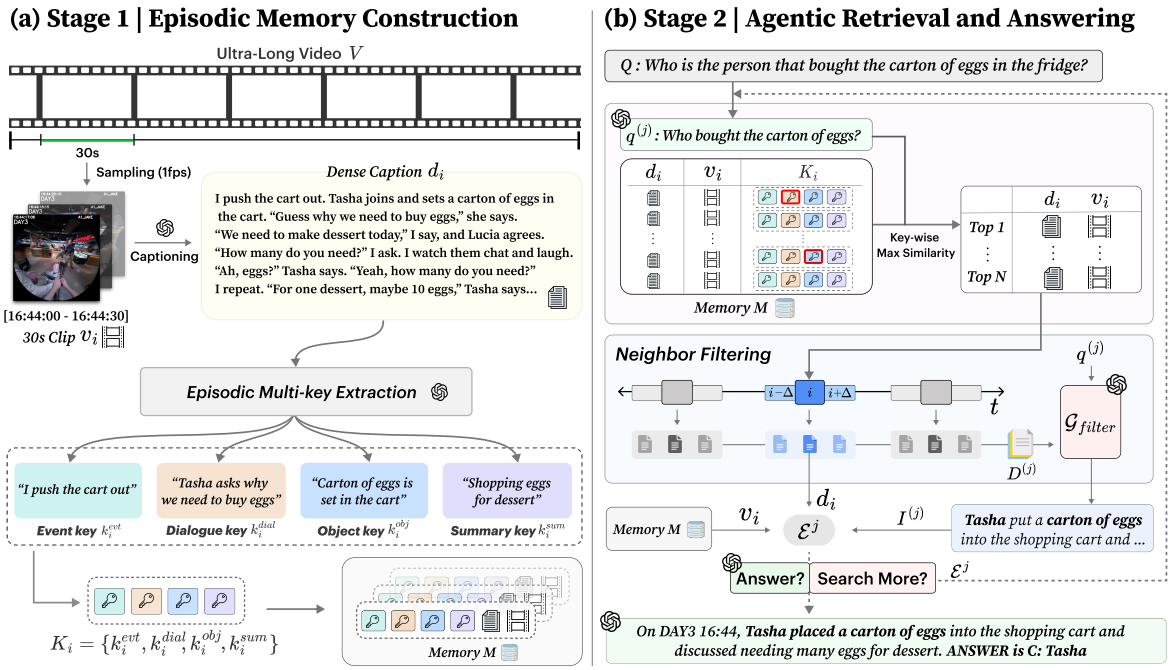


Figure 2. Overall Pipeline of MERIT. (a) **Stage 1: Episodic memory construction.** For each 30 seconds video clip, dense captions are generated. Subsequently, an episodic multi-key extraction process derives four distinct keys per clip, collectively forming the memory $\mathcal{M}$. (b) **Stage 2: Agentic retrieval and answering.** The solver regenerates a query and matches it against multi-key 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.

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 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, enabling clips to match the query via their 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 seconds clip. While short clips enable lightweight indexing, retrieving only the anchor 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 window

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

Table 1. Comparison of MERIT against various baselines on EgoLifeQA. Baseline results are from (Yeo et al., 2025; Rege et al., 2026). **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 (Bai et al., 2025a)	35.2	30.2	39.3	46.4	46.0	38.6
Gemini 2.5 Pro (Comanici et al., 2025)	43.2	40.5	41.0	55.2	52.4	46.4
GPT-5 (OpenAI, 2025)	47.2	42.1	47.5	53.6	55.6	48.6
Hierarchical Memory Based						
EgoRAG (Yang et al., 2025a) (GPT-5)	40.0	56.3	62.3	54.4	52.4	52.0
Ego-R1 (Yang et al., 2025b) (3B)	51.2	53.2	63.9	50.4	50.8	53.0
Graph Memory Based						
LightRAG (Guo et al., 2024) (GPT-5)	40.8	48.4	67.2	50.4	44.4	48.8
HippoRAG (Gutiérrez et al., 2025) (GPT-5)	48.8	60.3	70.5	60.8	66.7	59.6
Video-RAG (Luo et al., 2024) (GPT-5)	49.6	56.3	67.2	55.2	54.0	55.4
HippoMM (Lin et al., 2025) (GPT-5)	45.6	53.2	70.5	55.2	58.7	54.6
M3-Agent (Long et al., 2025) (7B)	44.4	54.8	62.3	56.8	54.0	53.5
EGAgent (Rege et al., 2026) (Gemini 2.5 Pro)	54.4	57.1	60.3	62.4	**74.6**	57.5
WorldMM (Yeo et al., 2025) (Qwen3-VL-8B)	49.6	56.4	63.9	58.4	58.7	56.4
WorldMM (Yeo et al., 2025) (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**

where Δ is an integer radius indicating the number of adjacent clips on each side (e.g., Δ=2 spans $\pm$1 minute under a 30 seconds clip setting). The full 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. EgoLifeQA (Yang et al., 2025a) is an ultra-long egocentric dataset averaging 44.3 hours per video, capturing six individuals over seven days with 500 QA pairs. EgoLifeQA features five QA types: "EntityLog" and "EventRecall" test precise factual grounding of objects and past events, while "HabitInsight", "RelationMap", and "TaskMaster" evaluate semantic reasoning over human behaviors, interactions, and future task planning. LVBench (Wang et al., 2025) assesses general long-form understanding, featuring 1,549 questions from 103 videos with an average duration of 4,101 seconds. Video-MME (Fu et al., 2025) (Long Subset) comprises videos ranging from 30 to 60 minutes, containing 900 questions uniformly allocated as three pairs per video. All tasks are formulated as multiple-choice questions (MCQs), with QA accuracy as the primary evaluation metric.

4.2. Implementation Details

To evaluate our framework, we applied our framework in Qwen3-VL-8B (Bai et al., 2025a), GPT-5 (OpenAI, 2025), and Gemini 2.5 Pro (Comanici et al., 2025). Following the previous work (Yeo et al., 2025), we employ GPT-5-mini to generate dense captions and episodic multi-key for each 30 seconds clip. For EgoLifeQA and Video-MME(Long), we incorporate ASR transcripts as an additional modality during captioning (Rege et al., 2026; Yeo et al., 2025). For LVBench, 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 supplementary material.

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 archi-

Table 2. Comparison of MERIT against various baselines on LVBench and Video-MME(L) benchmarks.

Model	LVBench	Video-MME(L)
MLLMs (Uniform Sampling)		
Gemini 2.5 Pro (Comanici et al., 2025)	57.0	55.7
GPT-5 (OpenAI, 2025)	60.4	74.3
Hierarchical Memory Based		
EgoRAG (Yang et al., 2025a) (GPT-5)	32.2	41.1
Ego-R1 (Yang et al., 2025b) (3B)	34.1	42.7
Graph Memory Based		
LightRAG (Guo et al., 2024) (GPT-5)	30.4	46.6
HippoRAG (Gutiérrez et al., 2025) (GPT-5)	54.0	52.1
Video-RAG (Luo et al., 2024) (GPT-5)	33.1	55.4
HippoMM (Lin et al., 2025) (GPT-5)	38.2	41.6
M3-Agent (Long et al., 2025) (7B)	49.3	55.3
EGAgent (Rege et al., 2026) (Gemini 2.5 Pro)	-	74.1
WorldMM (Yeo et al., 2025) (GPT-5)	61.9	76.6
Ours		
MERIT (Gemini 2.5 Pro)	-	76.3
MERIT (GPT-5)	**71.8**	**77.7**

tectures. 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, 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.

Table 3. Ablation study on Neighbor Filtering (NF) across two solvers. Target Time Oracle(Oracle) provides ground-truth temporal segments as an upper bound, while MERIT uses retrieved segments. Hit. denotes Hit Rate.

EXP	NF	Ent.	EvR.	Hab.	Rel.	Task	Avg	Hit.
Oracle (Qwen3-VL-8B)	✗	**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)	✗	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**
Oracle (GPT-5)	✗	**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)	✗	**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**

4.4. Ablation Studies

We conduct ablation studies to evaluate the individual components of MERIT. All ablation experiments are conducted on the EgoLifeQA 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.

4.4.1. 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 seconds clip, without any adjacent context.

To establish an upper bound for this single-clip setting, we evaluate a "Target Time Oracle(Oracle)" configuration, which directly provides the solver with frames and captions from the ground-truth 30 seconds 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.

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 seconds segment often lacks sufficient episodic context

Table 4. Performance analysis based on key combinations with Qwen3-VL-8B. 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

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.

4.4.2. Effect of Multi-Key Combinations

Table 4 details the performance variations across different configurations of our multi-key representation using the Qwen3-VL-8B(Bai et al., 2025a). 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.

Table 5. Retrieval accuracy comparison with baseline models using GPT-5. 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. 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	Ent.	EvR.	Hab.	Rel.	Task	Avg
Qwen3-VL-4B	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	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	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	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**

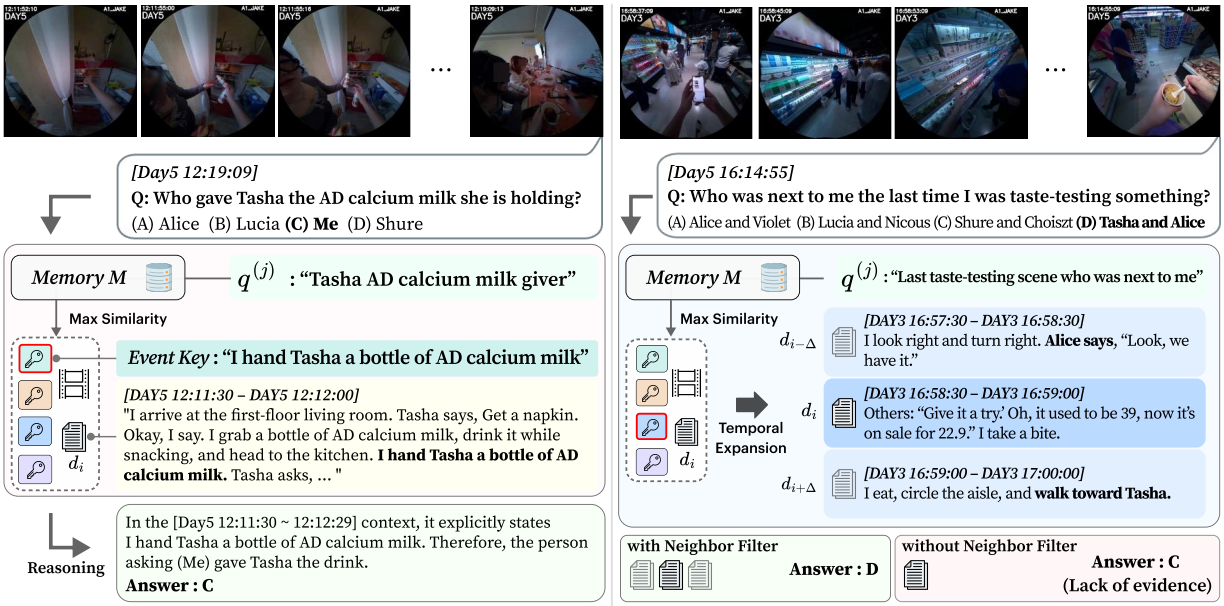
4.4.3. Retrieval Accuracy vs. Baselines

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).

Compared to the hierarchical memory of EgoRAG (Yang et al., 2025a) and the graph memory of WorldMM (Yeo et al., 2025), 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 seconds 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 seconds 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.

4.4.4. Effect of Retrieved Clip Quantity Across Solver Capacities

Table 6 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-



(a) Multi-Key Memory and Retrieval (b) Temporal Expansion via Neighbor Filtering

Figure 3. Qualitative results of MERIT on EgoLifeQA (Yang et al., 2025a). (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.

source models (Bai et al., 2025a), 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 (Comanici et al., 2025; OpenAI, 2025) demonstrate substantial improvements when provided with a larger pool of clips. Increasing to Top 10 boosts the accuracy of Gemini 2.5 Pro (Comanici et al., 2025) (61.4% to 64.2%) and GPT-5 (OpenAI, 2025) (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 the redundant information in additional temporal windows, they successfully 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 (Yang et al., 2025a). 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 directly to the correct answer.

Effect of Neighbor Filtering. Fig. 3 (b) presents a Relation-Map case from EgoLifeQA (Yang et al., 2025a), requiring an understanding of interpersonal relationships and interactions. The query asks about an event from two days before the query time, a scene 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, allowing the query-aware filter to extract the necessary 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.

References

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*, 2025a.

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., and Lin, J. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

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.

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.

Guo, Z., Xia, L., Yu, Y., Ao, T., and Huang, C. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*, 2(3), 2024.

Gutiérrez, B. J., Shu, Y., Qi, W., Zhou, S., and Su, Y. From rag to memory: Non-parametric continual learning for large language models. *arXiv preprint arXiv:2502.14802*, 2025.

Jeong, S., Kim, K., Baek, J., and Hwang, S. J. Videorag: Retrieval-augmented generation over video corpus. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21278–21298, 2025.

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.

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.

Long, L., He, Y., Ye, W., Pan, Y., Lin, Y., Li, H., Zhao, J., and Li, W. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. 2025.

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.

OpenAI. Gpt-5 system card, August 2025. URL https://openai.com/index/gpt-5-system-card/.

Rege, A., Sadhu, A., Li, Y., Li, K., Vinayak, R. K., Chai, Y., Lee, Y. J., and Kim, H. J. Agentic very long video understanding. 2026.

Ren, X., Xu, L., Xia, L., Wang, S., Yin, D., and Huang, C. Videorag: Retrieval-augmented generation with extreme long-context videos. *arXiv preprint arXiv:2502.01549*, 2025.

Shen, X., Zhang, W., Chen, J., and Elhoseiny, M. Vgent: Graph-based retrieval-reasoning-augmented generation for long video understanding, 2025.

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.

Xu, Z., Zhang, J., Wang, Q., and Liu, Y. E-vrag: Enhancing long video understanding with resource-efficient retrieval augmented generation, 2025.

Xue, Z., Zhang, J., Xie, X., Cai, Y., Liu, Y., Li, X., and Tao, D. Adavideorag: Omni-contextual adaptive retrieval-augmented efficient long video understanding. *arXiv preprint arXiv:2506.13589*, 2025.

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, 2025a.

Yang, J. et al. Ego-r1. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025b.

Yeo, W., Kim, K., Yoon, J., and Hwang, S. J. Worldmm: Dynamic multimodal memory agent for long video reasoning. *arXiv preprint arXiv:2512.02425*, 2025.

Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., and Li, C. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.