EM^{2}Mem: Event-Centric Multimodal Memory for Large Language Models

Yijun Chen[1*], **Yaqi Zheng**[1*], **Yanya Li**[1], **Boyi Xiao**[2],
Buqiang Xu[1], **Shuofei Qiao**[1], **Jizhan Fang**[1], **Xinle Deng**[1], **Yunzhi Yao**[1],
Xuehai Wang[1], **Liuxin Zhang**[3], **Hui Li**[3], **Huajun Chen**[1], **Shumin Deng**[1†]
[1]Zhejiang University [2]South China University of Technology
[3]Lenovo Group Limited
ie_yijunchen@zju.edu.cn, 231sm@zju.edu.cn

Abstract

Multimodal memory offers a scalable interface for long-video question answering, but existing methods often retrieve captions, frames, transcripts, summaries, or graph facts as isolated fragments. Although searchable, such fragments are not generation-ready: language models must reconstruct cross-modal and temporal alignments at inference time, when context is limited and attribution is difficult. We propose EM^{2}Mem, an event-centric multimodal memory framework that binds heterogeneous evidence to event anchors during memory construction. Each event-indexed memory cell aligns multimodal records, temporal context, graph-linked relations, semantic facts, and provenance, enabling compact evidence readout over grounded multimodal events rather than modality-specific fragments. Across three long-video QA benchmarks, EM^{2}Mem improves average accuracy over the strongest memory baseline by 2.0, 2.4, and 3.7 points, improves strict event-level Top-5 evidence recall by 7.0 points, and reduces per-query latency by $4.67\times$ and total inference tokens by 63.66%[1].

1 Introduction

Multimodal memory is becoming essential for enabling Large Language Models (LLMs) to reason over extended multimodal experiences. Unlike short-image or short-video understanding, long-video QA requires models to answer questions whose supporting evidence may appear sparsely across minutes or hours of visual observations, speech transcripts, OCR, scenes, objects, and recurring behaviors (Yang et al., 2025; Tian et al., 2025; Fu et al., 2025). Since such evidence cannot be reliably captured by a single context window or a flat video representation, LLMs need an external memory that can store observations, retrieve task-relevant context, and support grounded generation. The key problem, is not only how much information to store, but how to organize heterogeneous evidence into retrievable units that align with the way questions are asked and answers are justified.

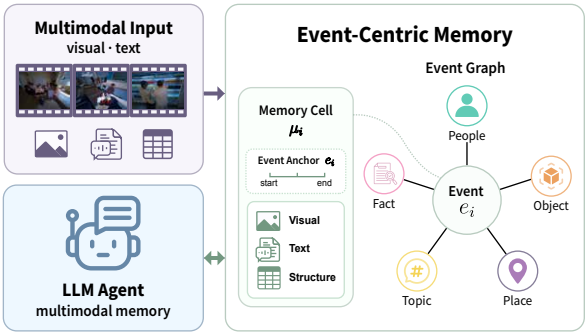


Figure 1: EM^{2}Mem organizes multimodal evidence into event-centric memory cells, enabling grounded retrieval over events rather than isolated fragments.

Recent multimodal memory and video retrieval-augmented generation methods typically construct fragment-centric memories by compressing videos into modality- or structure-specific stores, such as captions, summaries, keyframes, transcript segments, embeddings, triples, or knowledge graphs (Luo et al., 2024; Kahatapitiya et al., 2025; Ma et al., 2025; Gutierrez et al., 2024; Yeo et al., 2025). Although these fragments are searchable, they often fail to provide the event-level grounding needed for long-video QA. Questions commonly refer to events, situations, recurring behaviors, and temporally grounded relations, rather than isolated frames, sentences, or graph edges. As a result, existing systems face a *retrieve-then-align* bottleneck: they first retrieve disconnected fragments and then rely on the LLM to reconstruct missing cross-modal and temporal bindings during generation, when the context budget is limited and attribution is the most difficult to maintain.

To address this bottleneck, we rethink the re-

[*]Equal Contribution.
[†]Corresponding Author.
[1]The code will be integrated into https://github.com/zjunlp/LightMem.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 19117–19144
October 24-29, 2026 ©2026 Association for Computational Linguistics

trieval unit of multimodal memory: ***How should multimodal memory organize heterogeneous evidence for LLM agents?*** Inspired by event cognition, which views events as coherent units for organizing perception, language, actions, and temporal boundaries (Zacks et al., 2007; Radvansky and Zacks, 2014; Li et al., 2022b), we propose **EM^{2}Mem**, an **Event-Centric Multimodal Memory** framework for long-video QA. Rather than treating events as merely temporal video segments, EM^{2}Mem uses event anchors as language-addressable memory indices that bind captions, transcripts, keyframes, structured metadata, temporal context, and provenance during memory construction. This yields an *align-then-retrieve* design: heterogeneous evidence is unified into event-indexed multimodal memory cells before retrieval, so the model retrieves grounded event-level evidence rather than assembling isolated fragments afterward. To support reasoning beyond individual events, EM^{2}Mem further builds an episodic graph for cross-event relations and temporal transitions, together with a semantic graph for long-term facts, habits, preferences, and stable relations, each grounded in supporting event evidence.

This event-centric design yields consistent gains in accuracy, efficiency, and attribution. On three long-video QA benchmarks, EgoLifeQA, Ego-R1 Bench, and Video-MME (L), EM^{2}Mem improves average accuracy over the strongest memory baseline by 2.0%, 2.4%, and 3.7% respectively, while reducing per-query latency by 4.67 times and total inference tokens by 63.66%. It also localizes evidence more precisely, improving strict event-level Top-5 recall by 7.0 points over WorldMM's five rounds of iterative retrieval. Further analyses show that structured event-centric evidence provides a more effective memory interface than raw frames or flattened captions, and that construction-time unification outperforms retrieval-time fusion. Together, these results suggest that organizing multimodal memory around events provides a more natural and effective foundation for grounded, attributable long-video QA.

2 Background

Given a long video with visual and text-form evidence streams $V = \{V^{\text{vis}}, V^{\text{text}}\}$ and a question q, long-video QA aims to generate an answer $\hat{y}$ grounded in video-specific evidence. Here, V^{text} includes transcripts, OCR, subtitles, dialogue, and other textual cues (Luo et al., 2024; Cheng et al., 2024; Chen et al., 2023). Since relevant evidence is often sparse, temporally distant, and distributed across modalities, directly feeding the full video into an LLM or MLLM is inefficient and may obscure cross-modal dependencies.

Existing multimodal memory systems typically convert long videos or agent experiences into searchable modality-typed units (Yeo et al., 2025; Long et al., 2025; Fan et al., 2024; Liu et al., 2026):

$$\begin{aligned} u_j^m &= \left(s_j^m, e_j^m, p_j^m, \tau_j, m, \ell_j\right), \\ m &\in \{\text{vis}, \text{text}\}, \\ s_j^{\text{vis}} &= c_j^{\text{clip}} = \phi_{\text{cap}}(F_j, K_j), \\ s_j^{\text{text}} &= \phi_{\text{text}}(r_j, o_j, d_j). \end{aligned} \tag{1}$$

Here, s_j^m and e_j^m denote the searchable summary and embedding, p_j^m points to the source evidence, τ_j is the timestamp, and ℓ_j stores links to other units. Visual frames F_j and keyframes K_j are converted into clip captions, while transcripts r_j, OCR/subtitles o_j, and other textual cues d_j are normalized into text summaries. Although such units provide a unified storage interface, they remain organized around individual modality-specific signals or compressed observations rather than coherent multimodal events.

3 Method

3.1 Overview

We propose **EM^{2}Mem**, an **Event-Centric Multimodal Memory** framework for long-video reasoning with LLMs. EM^{2}Mem follows an *align-then-retrieve* design for long-video memory, as illustrated in Figure 2. It first parses a long video into short temporal events and uses event anchors as shared indices to align heterogeneous evidence during memory construction, rather than retrieving isolated modality-specific fragments at inference time. To support reasoning beyond individual events, EM^{2}Mem further builds lightweight event-linked graph memory: an episodic graph captures concrete cross-event relations, including shared entities, objects, scenes, topics, and temporal transitions, while a semantic graph stores higher-level regularities grounded in supporting events. At inference time, EM^{2}Mem retrieves relevant event cells, expands them with graph-linked evidence, and compiles a compact query-specific evidence view for answer generation.

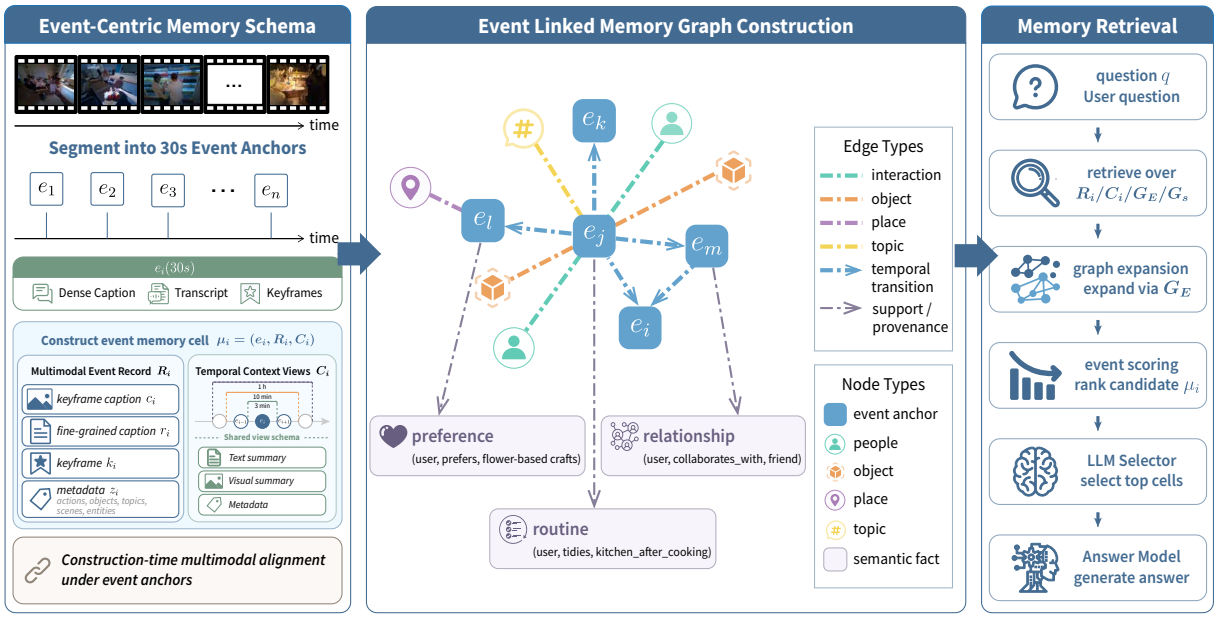


Figure 2: Overview of EM^{2}Mem. EM^{2}Mem parses a long video into event anchors, constructs event-centric multimodal memory cells with local records R_i and multi-scale context views $\mathcal{C}_i$, and links them through episodic and semantic graphs G_E and G_S. Given a question, lightweight retrieval selects and expands relevant event cells to compile evidence for answer generation.

3.2 Event-Centric Multimodal Memory Schema

Instead of treating a video as a flat sequence of frames or captions, EM^{2}Mem represents it as an event-indexed memory. We divide the video into short base segments, e.g. 30-second clips, and associate each segment with an event anchor:

$$e_i = (i, \tau_i^s, \tau_i^e)\,. \tag{2}$$

An event anchor is a temporal address rather than memory content. It provides a shared indexing key for heterogeneous evidence and higher-level abstractions. Formally, EM^{2}Mem represents the video as a set of **Event-Centric Multimodal Memory Cells**:

$$\begin{aligned} \mathcal{M}_{\text{EM}^2\text{Mem}} &= \left(\{\mu_i\}_{i=1}^{N}, G_E, G_S\right), \\ \mu_i &= (e_i, R_i, \mathcal{C}_i). \end{aligned} \tag{3}$$

Each event memory cell μ_i is anchored at e_i and stores local multimodal evidence R_i and multi-scale temporal context views $\mathcal{C}_i$. Beyond cell-level memory, EM^{2}Mem builds two global event-linked graphs: G_E captures cross-event episodic relations and temporal transitions, while G_S captures long-term semantic relations.

To make this schema concrete, suppose an event segment shows a person discussing a flower-based craft plan while interacting with flowers and vases. The corresponding memory cell binds the dialogue context, keyframe captions, and structured fields such as Act_i = *discussing a craft plan*, $\text{Obj}_i = \{\textit{flowers}, \textit{vases}\}$, and Top_i = *flower-based craft* under the same event anchor e_i, rather than storing them as separate captions, transcript snippets, and frames. Its context views $\mathcal{C}_i$ link the event to surrounding temporal summaries, while the graphs connect it to related entities, objects, topics, and recurring patterns. This design makes the memory cell query-ready: questions about the plan, involved objects, or nearby activities can retrieve one grounded event cell and its linked context instead of stitching together isolated modality-specific fragments at inference time. All evidence is grounded to event anchors via $\alpha(\cdot)$, preserving heterogeneous memory forms within a event-centric schema.

3.3 Multimodal Memory Construction

Multimodal Event Record. For each base event segment, EM^{2}Mem constructs a multimodal event record that stores the local evidence observed within the event:

$$\begin{aligned} R_i &= (c_i, r_i, k_i, z_i, \tau_i), \\ z_i &= (\text{Act}_i, \text{Obj}_i, \text{Top}_i, \text{Scen}_i, \text{Ent}_i). \end{aligned} \tag{4}$$

Here, c_i is a visual keyframe caption, r_i is the transcript or dialogue context, k_i denotes representative keyframes, and z_i contains structured metadata, including actions, objects, dialogue topics, visual scenes, and visual entities. These fields are

obtained through multimodal parsing modules, including captioning, transcript alignment, keyframe selection, and structured field extraction.

Temporal Context Views. To support reasoning at different temporal granularities, EM^{2}Mem further builds temporal context views over multiple scales, such as 3 minutes, 10 minutes, and 1 hour. At each scale, consecutive event anchors are grouped into a temporal block and summarized into a context view:

$$h_j^{(l)} = (u_{j,l}^{\text{text}}, u_{j,l}^{\text{vis}}, m_{j,l}, \tau_{j,l}, l). \tag{5}$$

The textual summary $u_{j,l}^{\text{text}}$ captures narrative and dialogue context, the visual summary $u_{j,l}^{\text{vis}}$ summarizes visual evidence, and $m_{j,l}$ aggregates normalized metadata such as actions, objects, scenes, and topics. Each context view is anchored to all base events covered by its temporal span, and the context-view set $\mathcal{C}_i$ of an event cell contains all views whose spans include e_i.

3.4 Event-Linked Memory Graph Construction

Beyond event memory cells, EM^{2}Mem builds two lightweight graph indices over shared event anchors: an episodic graph G_E and a semantic graph G_S. These graphs are used as auxiliary indices rather than as new graph formalisms. The episodic graph connects event anchors with typed entities such as people, objects, locations, scenes, and topics, and also includes temporal transitions between adjacent events. Each relation is grounded to its supporting event anchors, allowing graph-derived clues to be traced back to concrete video evidence.

The semantic graph summarizes recurring patterns from episodic evidence, such as habits, preferences, routines, and stable relationships. Unlike local event records, which describe concrete observations, semantic facts capture long-term regularities that may span multiple distant events. These facts are also linked back to their supporting anchors, so that high-level knowledge remains evidence-grounded. Together, G_E and G_S complement event memory cells by providing cross-event connectivity and long-term semantic context.

3.5 Memory Retrieval and Ranking

At inference time, EM^{2}Mem performs lightweight event-level readout over pre-aligned memory cells, which shifts retrieval from query-time cross-modal fusion to compact event-cell selection and expansion. Given a question q, it first identifies relevant event memory cells using signals from local multimodal records, multi-scale temporal context views, and graph-linked evidence. It then performs lightweight expansion through temporal transitions, shared entities, objects, locations, topics, and relevant semantic facts, so that complementary evidence from nearby or semantically related events can be included.

An LLM-based selector further filters the candidate event cells and keeps only the evidence needed to answer the question. For each selected event, EM^{2}Mem compiles a query-specific evidence view E_q, including relevant captions, transcripts, structured visual fields, temporal summaries, semantic facts, and selected keyframes. The final answer $\hat{y}$ is generated by LLM_{ans} conditioned on the question q and the compact evidence view E_q.

4 Experiment

4.1 Experimental Setup

Datasets and Metrics. We evaluate EM^{2}Mem on three multiple-choice long-video QA benchmarks: EgoLifeQA, Ego-R1 Bench, and Video-MME (L) (Yang et al., 2025; Tian et al., 2025; Fu et al., 2025). Together, they cover week-long egocentric daily-life QA, ultra-long egocentric reasoning, and open-domain long-video understanding. Following the standard protocol of each benchmark, we report category-level and average accuracy. Dataset statistics, category definitions, and license details are provided in Appendix A.

Baselines. We compare EM^{2}Mem with representative baselines spanning base MLLMs, long-video MLLMs, RAG-based video QA methods, and memory-based long-video reasoning frameworks, as listed in Tables 1 and 3. For the strongest and most related memory baseline, we report both the originally published WorldMM results and our reproduced WorldMM† results under the same evaluation setting as EM^{2}Mem. Detailed baseline configurations are provided in Appendix B.1.

4.2 Main Results

Accuracy across long-video QA benchmarks. Tables 1 and 3 report the main results on EgoLifeQA, Ego-R1 Bench, and Video-MME (L). Overall, EM^{2}Mem achieves strong performance against base MLLMs, long-video LLMs, RAG-based methods, and memory-based video reason-

Model	EgoLifeQA						Ego-R1 Bench					
	Ent.	EvR.	Hab.	Rel.	Task	Avg.	Ent.	EvR.	Hab.	Rel.	Task	Avg.
Base Models												
Qwen3-VL-8B	35.2	30.2	39.3	46.4	46.0	38.6	31.8	41.5	38.5	42.1	44.7	35.7
Gemini 2.5 Pro	43.2	40.5	41.0	55.2	52.4	46.4	43.9	56.1	53.9	47.4	47.4	46.7
GPT-5	47.2	42.1	47.5	53.6	55.6	48.6	41.8	58.5	53.9	52.6	50.0	46.3
Long Video LLMs												
VideoChat-Flash	28.8	32.5	37.7	37.6	38.1	34.2	43.4	43.9	38.5	31.6	44.7	42.7
Time-R1	39.2	50.8	65.6	48.8	47.6	48.8	49.2	48.8	46.2	42.1	44.7	48.0
Video-RTS	40.8	48.4	62.3	48.8	47.6	48.2	47.6	46.3	53.9	52.6	47.4	48.0
RAG-based Video LLMs												
LightRAG	40.8	48.4	67.2	50.4	44.4	48.8	54.0	61.0	46.2	42.1	42.1	52.3
HippoRAG	48.8	60.3	70.5	60.8	66.7	59.6	54.5	65.9	69.2	52.6	50.0	56.0
Video-RAG	49.6	56.3	67.2	55.2	54.0	55.4	48.7	58.5	53.9	47.4	44.7	49.7
Memory-based Video LLMs												
EgoRAG	40.0	56.3	62.3	54.4	52.4	52.0	46.6	56.1	46.2	47.4	55.3	49.0
Ego-R1	51.2	53.2	63.9	50.4	50.8	53.0	50.8	63.4	38.5	36.8	57.9	52.0
HippoMM	45.6	53.2	70.5	55.2	58.7	54.6	51.9	56.1	46.2	52.6	57.9	53.0
M3-Agent	44.4	54.8	62.3	56.8	54.0	53.5	52.4	58.5	38.5	42.1	52.6	52.0
WorldMM	**62.4**	64.3	**75.4**	62.4	71.4	65.6	64.6	**70.7**	**76.9**	**57.9**	**63.2**	65.3
WorldMM†	57.6	**65.1**	68.9	68.8	60.3	64.0			–			
EM^{2}Mem *(Ours)*	60.8	61.1	63.9	**72.8**	**74.6**	**66.0**	**74.6**	53.7	69.2	47.4	57.9	**67.7**

Table 1: Performance comparison on EgoLifeQA and Ego-R1 Bench. WorldMM denotes the originally reported results. † denotes our reproduced results under the same evaluation setting as EM^{2}Mem. For Ego-R1 Bench, we report the published WorldMM results because reproducing WorldMM requires rebuilding subject-level long-video memories for all six participants, which incurs substantial construction and inference cost. The corresponding entries are marked as "–".

Metric	WorldMM	EM^{2}Mem	Gain
Per-query inference efficiency			
Avg. latency / query (s)	459.00	**98.21**	**4.67**×
End-to-end evaluation throughput			
Wall-clock time (s)	229,502	**6,138**	**37.39**×
Inference token cost			
Input tokens	31.95M	**13.29M**	**58.42%**↓
Output tokens	10.08M	**1.99M**	**80.29%**↓
Total tokens	42.03M	**15.27M**	**63.66%**↓

Table 2: Inference efficiency comparison on EgoLifeQA. Latency is measured per query; wall-clock time follows each method's practical evaluation pipeline.

ing systems. We report both previously published WorldMM results and our reproduced WorldMM† results for reference; since reproduced results are obtained under the same evaluation setting as EM^{2}Mem, they provide the most controlled comparison. Under this setting, EM^{2}Mem improves over WorldMM† by 2.0 points on EgoLifeQA (66.0 vs. 64.0), and 3.7 points on Video-MME (L) (76.8 vs. 73.1). On Ego-R1 Bench, where we use the originally reported WorldMM result for comparison, EM^{2}Mem achieves 67.7 average accuracy, outperforming WorldMM by 2.4 points (67.7 vs. 65.3). These gains are especially visible on RelationMap and TaskMaster in EgoLifeQA, EntityLog in Ego-R1 Bench, and AREC, CNT, OCR, and ORES in Video-MME (L), suggesting that event-indexed retrieval units are useful for grounded entity tracking, task-state reasoning, and cross-modal evidence aggregation. Compared with the originally reported WorldMM numbers, EM^{2}Mem remains competitive and slightly higher on average, although WorldMM is stronger on several habit, temporal, and synthetic reasoning categories.

Inference efficiency and scalability. Beyond accuracy, Table 2 compares EM^{2}Mem with WorldMM along three efficiency dimensions: per-query latency, end-to-end wall-clock time, and inference token cost. We report both latency and wall-clock time because they capture different aspects of efficiency: latency measures the delay of an individual request, while wall-clock time reflects end-to-end throughput under each method's practical execution pipeline. In this setting, EM^{2}Mem processes independent questions with 8 workers, whereas WorldMM is constrained by incremental HippoRAG graph construction during inference. EM^{2}Mem reduces average latency from 459.00s to 98.21s, yielding a $4.67\times$ per-query speedup. This

Model	ARES	AREC	ATTR	CNT	ISYN	OCR	ORES	OREC	SPER	SRES	TPER	TRES	Avg.
Base Models													
Qwen3-VL-8B	62.2	54.0	51.9	43.8	68.1	42.9	62.9	57.4	33.3	45.5	33.3	67.0	61.0
Gemini 2.5 Pro	56.9	47.6	66.7	41.7	71.8	57.1	53.3	40.7	0.0	72.7	**66.7**	48.4	55.7
GPT-5	71.1	69.8	70.4	47.9	**88.3**	57.1	75.8	74.1	33.3	72.7	50.0	75.8	74.3
Long Video LLMs													
VideoChat-Flash	35.0	42.9	37.0	31.3	34.4	42.9	60.0	46.3	33.3	54.5	33.3	46.2	44.1
Time-R1	20.6	28.6	25.9	35.4	31.9	35.7	53.3	48.2	33.3	36.4	50.0	44.0	37.6
Video-RTS	43.3	52.4	40.7	39.6	33.7	42.9	60.8	53.7	33.3	45.5	50.0	49.5	47.9
RAG-based Video LLMs													
LightRAG	41.7	30.2	40.7	35.4	54.0	50.0	46.7	61.1	33.3	45.5	50.0	52.8	46.6
HippoRAG	45.6	47.6	40.7	37.5	52.2	42.9	52.9	64.8	66.7	54.5	50.0	70.3	52.1
Video-RAG	51.7	47.6	37.0	39.6	49.7	57.1	62.1	68.5	66.7	45.5	50.0	68.1	55.4
Memory-based Video LLMs													
EgoRAG	31.1	55.6	33.3	22.9	41.1	28.6	44.6	48.2	33.3	54.5	**66.7**	48.4	41.1
Ego-R1	37.2	52.4	40.7	35.4	38.0	35.7	42.1	51.9	**66.7**	63.6	50.0	52.8	42.7
HippoMM	41.1	42.9	55.6	35.4	38.7	35.7	37.9	53.7	33.3	54.5	50.0	47.3	41.6
M3-Agent	52.2	57.1	59.3	45.8	51.5	42.9	54.6	64.8	33.3	45.5	50.0	71.4	55.3
WorldMM	**81.1**	73.0	70.4	54.2	85.3	42.9	75.0	**77.8**	33.3	72.7	**66.7**	**79.1**	76.6
WorldMM†	73.3	68.3	**77.8**	60.4	80.2	50.0	72.4	**77.8**	33.3	**90.9**	**66.7**	71.1	73.1
EM^{2}Mem *(Ours)*	77.2	**76.2**	**77.8**	**64.6**	80.7	**64.3**	**77.0**	**77.8**	33.3	81.8	50.0	**79.1**	**76.8**

Table 3: Category-wise performance on Video-MME (L). WorldMM denotes the originally reported results. † denotes our reproduced results under the same evaluation setting as EM^{2}Mem.

shows that the efficiency gain is not only due to parallel execution. The main reason is that EM^{2}Mem reads from pre-constructed event-indexed memory cells instead of performing evidence alignment and graph organization during inference. EM^{2}Mem further reduces wall-clock evaluation time from 229,502s to 6,138s, corresponding to a $37.39\times$ throughput improvement, and lowers total inference token cost from 42.03M to 15.27M, a 63.66% reduction. These results indicate that construction-time memory organization improves not only answer quality, but also the inference-time cost of evidence readout. Figure 3e further contextualizes these savings by amortizing the offline construction cost over repeated queries on the same memory. Although EM^{2}Mem incurs a higher upfront construction cost, its lower inference-time wall-clock cost reaches break-even after roughly 23–24 queries, while token usage requires a larger reuse scale to break even; we provide the full end-to-end derivation in Appendix C.2. Taken together, these results suggest that event-indexed retrieval units improve long-video QA accuracy while making inference-time evidence readout more scalable.

4.3 Ablation Study

Table 4 and Figure 3b show the overall ablation results on EgoLifeQA. Removing temporal context views leads to the largest drop, decreasing accuracy from 66.0% to 60.4%. Removing semantic facts and episodic graph support also substantially hurts performance, reducing accuracy to 61.4% and 61.6%, respectively. These results validate the design of combining local event records, temporal context views, episodic graph, and semantic facts in an event-indexed memory.

Method	Overall Acc.	Δ
EM^{2}Mem	**66.0**	–
w/o Temporal Context Views	60.4	-5.6
w/o Semantic Memory	61.4	-4.6
w/o Episodic Graph	61.6	-4.4
Local 30s Event Record	64.0	-2.0

Table 4: Overall ablation results on EgoLifeQA. Δ denotes the absolute accuracy drop compared with the full EM^{2}Mem model.

Interestingly, the local 30-second record retrieval baseline remains competitive, achieving 64.0%. This suggests that fine-grained local event record already provide strong evidence for short-range factual questions. However, the full EM^{2}Mem still achieves the best overall accuracy. The category-wise results in Table 11 at Appendix C.3 show that the advantage of EM^{2}Mem is more pronounced on relation- and task-oriented questions, where evidence often needs to be aggregated across temporally distant events and connected through entity, object, or activity relations.

4.4 Analysis

Event binding turns heterogeneous evidence into answerable memory units. We analyze how the representation and fusion stage of multimodal evidence affect memory-based long-video QA, using the first 250 questions of EgoLifeQA. Our analysis spans two dimensions: (1) memory representation of visual evidence, including raw frames, caption-level descriptions, and structured event fields; and (2) unification stage, comparing retrieval-time fusion, where independent stores are combined during inference, with construction-time unification, where cross-modal evidence is aligned under shared event anchors before indexing.

Structured event fields bridge multimodal evidence and natural-language questions. As shown in Table 5 and Figure 3a, the representation of visual evidence has a substantial impact on QA performance. Across both fusion stages, structured evidence outperforms raw and caption-level memories, achieving 71.2% accuracy when combined with construction-time unification. This suggests typed visual fields form a more effective memory interface for LLM-based QA: compared with raw frames, they are easier to retrieve, verbalize, and attribute; compared with flattened captions, they preserve explicit entities, actions, scenes, and other question-relevant cues. Rather than serving only as a visual representation, structured evidence acts as an answerable interface between multimodal observations and natural-language questions.

Memory Representation	Unification Stage	
	Retrieval	Construction
Raw frames	66.8	68.0
Flattened captions	65.2	65.6
Structured event fields	68.0	**71.2**

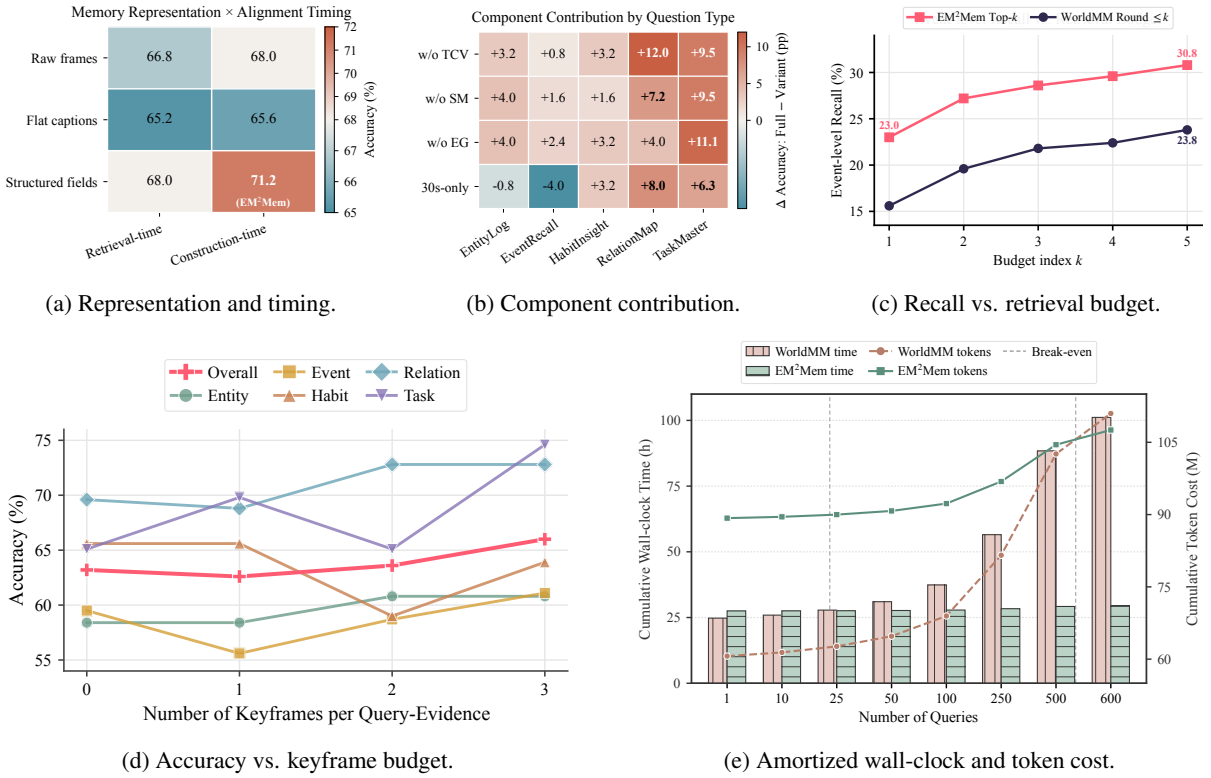
Table 5: Accuracy (%) on first 250 EgoLifeQA questions. Rows vary memory representation visual evidence; columns vary with unified multimodal evidence.

Construction-time unification mitigates retrieve-then-align bottlenecks. Table 5 and Figure 3a also show that construction-time unification consistently improves over retrieval-time fusion. Retrieval-time fusion follows a retrieve-then-align pattern: each modality is retrieved independently, so evidence that is weakly relevant in isolation may be discarded before it can be combined with complementary cues from other modalities. In contrast, construction-time unification aligns captions, transcripts, and structured visual fields into shared event-level retrieval units before indexing. This makes retrieved evidence more semantically coherent, temporally grounded, and generation-ready. It also preserves evidence whose relevance emerges only after cross-modal binding, such as an object mentioned visually but disambiguated by nearby dialogue, or an action described in text but grounded by keyframes. The gain is largest for structured evidence (+3.2%), suggesting that explicit fields such as entities and actions provide useful anchors for cross-modal alignment. Together, these results support our central claim: effective multimodal memory should unify heterogeneous evidence during memory construction, producing event-level retrieval units that are directly usable for grounded and attributable long-video QA.

Keyframes serve as lightweight verification after memory retrieval. We further study how many keyframes should be included in the query-specific evidence view after relevant event cells have been selected. This analysis asks whether event-level textual and structured memory is sufficient, or whether answer generation still benefits from a small amount of visual evidence for verification. As shown in Figure 3d, overall accuracy improves from 63.2% without keyframes to 66.0% with three keyframes. The gains are larger for TaskMaster and RelationMap, where visual evidence helps verify task states, participant identities, and interaction context. By contrast, HabitInsight benefits less consistently because it depends more on repeated long-term behavioral patterns than on local visual verification. Thus, EM^{2}Mem uses compact event-centric multimodal memory as primary evidence and keyframes only for lightweight grounding.

Type	WorldMM		EM^{2}Mem		Δ
	1R	5R	Top-1	Top-5	
Overall	15.6	23.8	23.0	**30.8**	+7.0
Ent.	13.6	21.6	20.0	**27.2**	+5.6
EvR.	19.8	**24.6**	19.0	23.8	-0.8
Hab.	8.2	23.0	27.9	**39.3**	+16.3
Rel.	16.0	24.0	21.6	**31.2**	+7.2
Task.	17.5	27.0	34.9	**42.9**	+15.9

Table 6: Strict 30-second event-level evidence recall (%). Δ denotes the absolute gain of EM^{2}Mem Top-5 over WorldMM 5R.



(a) Representation and timing.

(b) Component contribution.

(c) Recall vs. retrieval budget.

(d) Accuracy vs. keyframe budget.

(e) Amortized wall-clock and token cost.

Figure 3: Analysis of EM^{2}Mem design choices and efficiency. **(a)** Structured event fields with construction-time unification achieve the best accuracy. **(b)** Component-wise ablations show different contributions across question types. **(c)** EM^{2}Mem improves event-level recall as retrieval budget increases. **(d)** Visual keyframes improve overall accuracy, especially with three keyframes. **(e)** EM^{2}Mem amortizes its wall-clock overhead as the number of queries increases; dashed lines indicate wall-clock and token break-even points.

Event anchors enable single-pass evidence localization. We evaluate strict 30-second event-level recall, where a retrieval is correct only if the retrieved event anchor matches the annotated evidence segment. As shown in Table 6 and Figure 3c, EM^{2}Mem achieves 30.8% Top-5 recall, outperforming WorldMM after five iterative retrieval rounds by 7.0 points. Its Top-1 recall also reaches 23.0%, close to WorldMM 5R at 23.8%, indicating that event-indexed memory can localize relevant evidence with a single-pass retrieval procedure. The gains are largest on HabitInsight and TaskMaster, where evidence often depends on long-term patterns or task states, while EventRecall remains slightly lower, suggesting that localized recall questions may still benefit from iterative caption-level retrieval. We provide additional analysis of selector budgets and a qualitative case study in Appendix C.6 and Appendix C.8.

5 Related Work

Long Video Understanding. Video-language research has progressed from video-text pretraining and temporal-aware representation learning to video LLMs that connect visual encoders with LLMs for instruction following and long-form understanding (Sun et al., 2019; Tong et al., 2022; Wang et al., 2024b; Zhang et al., 2023; Maaz et al., 2024a; Jin et al., 2024; Li et al., 2025a). Recent methods extend video LLMs to longer contexts through visual token compression, sparse or hierarchical memory, long-context adaptation, and temporal reasoning mechanisms (Zhang et al., 2025b; Shen et al., 2025; Li et al., 2025b; Ren et al., 2024; Wang et al., 2025b,c,a). However, as videos scale from minutes to hours or days, relevant evidence becomes sparse, temporally distant, and distributed across modalities, motivating explicit mechanisms for organizing and retrieving video information.

Multimodal Memory for LLMs. Memory- and retrieval-augmented systems address long-video QA by storing captions, transcripts, keyframes, OCR, clip embeddings, or graph-structured indices and retrieving question-relevant evidence before generation (Guo et al., 2025; Gutierrez et al., 2024; Luo et al., 2024; Jeong et al., 2025; Kahatapitiya

et al., 2025). Closely related egocentric and agentic systems construct hierarchical memories, episode summaries, or adaptive retrieval procedures for long-range reasoning (Yang et al., 2025; Lin et al., 2025; Fan et al., 2024; Long et al., 2025; Tian et al., 2025; Yeo et al., 2025; Yin et al., 2025). We provide an extended related work in Appendix D.

6 Conclusion

We presented EM^{2}Mem, an event-centric multimodal memory framework that organizes heterogeneous evidence under shared event anchors. By retrieving aligned multimodal events rather than isolated fragments, EM^{2}Mem improves accuracy, evidence localization, and inference efficiency for scalable long-video question answering.

Limitations

EM^{2}Mem has several limitations. Its structured memory design trades visual fidelity for searchability: converting visual evidence into textual fields makes it compatible with event-level retrieval, but may lose fine-grained pixel details such as small objects, colors, layouts, and subtle visual states. Future work may explore coarse-to-fine multimodal memory, where structured fields support retrieval while finer visual evidence is preserved for verification. The align-then-retrieve paradigm is also not fully complete. Although multimodal evidence is aligned under event anchors before retrieval, the final answer stage still relies on selected keyframes for visually detailed reasoning. This leaves part of cross-modal alignment to inference time and may introduce modality bias. EM^{2}Mem also depends on upstream MLLMs and visual tools, so errors in captioning, object extraction, or metadata normalization may propagate into event cells and graphs. In addition, the framework shifts computation from inference to memory construction, making it more suitable for videos processed once and queried repeatedly than for real-time scenarios.

Ethical Statement

This work studies long-video question answering with multimodal memory, including evaluations on egocentric daily-life benchmarks. Such videos may contain privacy-sensitive information, including personal routines, indoor environments, object usage, social interactions, and potentially identifying visual or textual cues. We do not collect new human-subject data, and all experiments are conducted on publicly released benchmark datasets under their intended research settings. We do not attempt to identify individuals, infer protected attributes, or disclose raw video content or personally identifiable information; all results are reported in aggregate. Since EM^{2}Mem organizes events, entities, routines, and relations into structured memory, misuse on unconstrained personal videos could increase risks of privacy leakage, behavioral profiling, or surveillance-like applications. Therefore, practical deployment of such memory systems should require informed consent, access control, data minimization, memory deletion or expiration mechanisms, and safeguards against using inferred semantic facts for high-stakes decisions. We also note that egocentric video benchmarks may reflect demographic and environmental biases, so performance should not be assumed to generalize uniformly across populations or deployment contexts.

Acknowledgements

We would like to express our sincere gratitude to the anonymous reviewers for their thoughtful and constructive feedback. This work was supported by the New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0122802), National Natural Science Foundation of China (No.62676362, No.NSFCU23B2055, No.NSFCU19B2027), the Fundamental Research Funds for the Central Universities (226-2023-00138), the Yongjiang Talent Introduction Programme (2021A-156-G), and the Information Technology Center and State Key Lab of CAD&CG, Zhejiang University. This work was sponsored by CCF-Lenovo Blue Ocean Research Fund.

References

Pengfei Cao, Yuheng Chen, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. 2026. One mind, many tongues: A deep dive into language-agnostic knowledge neurons in large language models. *Artif. Intell.*, 353:104490.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Lin Bin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. 2024. Sharegpt4video: Improving video understanding and generation with better captions. 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.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2023. VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset. 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*.

Tao Chen, Shaobo Ju, Qiong Wu, Chenxin Fang, Kun Zhang, Jun Peng, Hui Li, Yiyi Zhou, and Rongrong Ji. 2025. Towards effective and efficient long video understanding of multimodal large language models via one-shot clip retrieval. *CoRR*, abs/2512.08410.

Yijun Chen, Boyi Xiao, Yixian Zhao, Haoting Xia, Buqiang Xu, Jizhan Fang, Yanya Li, Yaqi Zheng, Xuehai Wang, Zirui Xue, and 1 others. 2026a. Lightmem-ego: Your ai memory for everyday life. *arXiv preprint arXiv:2607.11487*.

Zhuoen Chen, Dongfang Li, Meishan Zhang, Baotian Hu, and Min Zhang. 2026b. Dynamic long context reasoning over compressed memory via end-to-end reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8064–8083.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *CoRR*, abs/2406.07476.

Xinle Deng, Yida Xue, Yijun Chen, Mingjun Mao, Ruobin Zhong, Buqiang Xu, Jizhan Fang, Haoming Xu, Tingwei Wu, Yajing Xu, and 1 others. 2026a. Mobilemem: Evaluating long-horizon memory for language agents in real-world mobile environments. In *ICLR 2026 Workshop on Lifelong Agents: Learning, Aligning, Evolving*.

Xinle Deng, Yida Xue, Xiangyuan Ru, Haoming Xu, Shuofei Qiao, Mengru Wang, Yijun Chen, Buqiang Xu, Chen Jiang, Yuchen Eleanor Jiang, and 1 others. 2026b. Mobilemem: Learning from a year of mobile experiences. *arXiv preprint arXiv:2608.13606*.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. 2024. Videoagent: A memory-augmented multimodal agent for video understanding. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXII*, Lecture Notes in Computer Science, pages 75–92. Springer.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. 2025. Lightmem: Lightweight and efficient memory-augmented generation. *CoRR*, abs/2510.18866.

Chaoyou Fu, Yuhan Dai, Yongdong 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, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 24108–24118. Computer Vision Foundation / IEEE.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2025. Lightrag: Simple and fast retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 10746–10761. Association for Computational Linguistics.

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. 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*.

Seungju Han, Jack Hessel, Nouha Dziri, Yejin Choi, and Youngjae Yu. 2023. Champagne: Learning real-world conversation from large-scale web videos. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 15452–15463. IEEE.

Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. 2025. Videorag: Retrieval-augmented generation over video corpus. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 21278–21298. Association for Computational Linguistics.

Peng Jin, Jinfa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David A. Clifton, and Jie Chen. 2022. Expectation-maximization contrastive learning for compact video-and-language representations. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13700–13710. IEEE.

Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S. Ryoo. 2025. Language repository for long video understanding. In *Findings of the Association for Computational Linguistics, ACL*

2025, Vienna, Austria, July 27 - August 1, 2025, Findings of ACL, pages 5627–5646. Association for Computational Linguistics.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025a. Llava-onevision: Easy visual task transfer. *Trans. Mach. Learn. Res.*, 2025.

Dongfang Li, Zixuan Liu, Gang Lin, Baotian Hu, and Min Zhang. 2026a. Lycheecluster: Efficient long-context inference with structure-aware chunking and hierarchical KV indexing. In *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pages 7607–7623. Association for Computational Linguistics.

Dongfang Li, Zixuan Liu, Junmai Wang, Jiahe Huang, Fuhao Li, Bonian Jia, Baotian Hu, and Min Zhang. 2026b. Lycheememory v2: Efficient long-term memory for llm agents via semantic segment-level consolidation. *Preprint*, arXiv:2608.12990.

Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven C. H. Hoi. 2022a. Align and prompt: Video-and-language pre-training with entity prompts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 4943–4953. IEEE.

Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. 2022b. Clip-event: Connecting text and images with event structures. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16399–16408. IEEE.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. 2025b. Videochat-flash: Hierarchical compression for long-context video modeling. *CoRR*, abs/2501.00574.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024. Llama-vid: An image is worth 2 tokens in large language models. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVI*, Lecture Notes in Computer Science, pages 323–340. Springer.

Niu Lian, Yuting Wang, Hanshu Yao, Jinpeng Wang, Bin Chen, Yaowei Wang, Min Zhang, and Shu-Tao Xia. 2026. From verbatim to gist: Distilling pyramidal multimodal memory via semantic information bottleneck for long-horizon video agents. *CoRR*, abs/2603.01455.

Zhijia Liang, Jiaming Li, Weikai Chen, Yanhao Zhang, Haonan Lu, and Guanbin Li. 2026. Oasis: On-demand hierarchical event memory for streaming video reasoning. *arXiv preprint arXiv:2604.17052*.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5971–5984. Association for Computational Linguistics.

Yueqian Lin, Qinsi Wang, Hancheng Ye, Yuzhe Fu, Hai Helen Li, and Yiran Chen. 2025. Hippomm: Hippocampal-inspired multimodal memory for long audiovisual event understanding. *CoRR*, abs/2504.10739.

Jiaqi Liu, Zipeng Ling, Shi Qiu, Yanqing Liu, Siwei Han, Peng Xia, Haoqin Tu, Zeyu Zheng, Cihang Xie, Charles Fleming, Mingyu Ding, and Huaxiu Yao. 2026. Omni-simplemem: Autoresearch-guided discovery of lifelong multimodal agent memory. *CoRR*, abs/2604.01007.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. 2025. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. *CoRR*, abs/2508.09736.

Yongdong Luo, Xiawu Zheng, Xiao Yang, Guilin Li, Haojia Lin, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. 2024. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *CoRR*, abs/2411.13093.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. 2025. Drvideo: Document retrieval based long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 18936–18946. Computer Vision Foundation / IEEE.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. 2024a. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 12585–12602. Association for Computational Linguistics.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024b. Videogpt+: Integrating image and video encoders for enhanced video understanding. *CoRR*, abs/2406.09418.

Chang Nie, Chaoyou Fu, Yifan Zhang, Haihua Yang, and Caifeng Shan. 2026. Personavlm: Long-term personalized multimodal llms. *CoRR*, abs/2604.13074.

OpenAI. 2026. Openai GPT-5 system card. *CoRR*, abs/2601.03267.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang,

and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5368–5393. Association for Computational Linguistics.

Gabriel A Radvansky and Jeffrey M Zacks. 2014. *Event cognition*. Oxford University Press.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14313–14323. IEEE.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. 2025. Longvu: Spatiotemporal adaptive compression for long video-language understanding. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 18221–18232. IEEE.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7463–7472. IEEE.

Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu. 2022. Long-form video-language pre-training with multimodal temporal contrastive learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Gemini Team. 2025a. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261.

Qwen Team. 2025b. Qwen3-vl technical report. *CoRR*, abs/2511.21631.

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang 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.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Jiawei Wang, Liping Yuan, and Yuchen Zhang. 2024a. Tarsier: Recipes for training and evaluating large video description models. *CoRR*, abs/2407.00634.

Shijian Wang, Jiarui Jin, Xingjian Wang, Linxin Song, Runhao Fu, Hecheng Wang, Zongyuan Ge, Yuan Lu, and Xuelian Cheng. 2025a. Video-thinker: Sparking "thinking with videos" via reinforcement learning. *CoRR*, abs/2510.23473.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, and 1 others. 2025b. Time-r1: Post-training large vision language model for temporal video grounding. *arXiv preprint arXiv:2503.13377*.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. 2024b. Internvideo2: Scaling foundation models for multimodal video understanding. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXV*, Lecture Notes in Computer Science, pages 396–416. Springer.

Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and Mohit Bansal. 2025c. Video-rts: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 28126–28140. Association for Computational Linguistics.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2025d. Videotree: Adaptive tree-based video representation for LLM reasoning on long videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 3272–3283. Computer Vision Foundation / IEEE.

Siwei Wen, Zhangcheng Wang, Xingjian Zhang, Lei Huang, and Wenjun Wu. 2026. Eventmemagent:

Hierarchical event-centric memory for online video understanding with adaptive tool use. *CoRR*, abs/2602.15329.

Buqiang Xu, Yijun Chen, Jizhan Fang, Ruobin Zhong, Yunzhi Yao, Yuqi Zhu, Lun Du, and Shumin Deng. 2026a. Structmem: Structured memory for long-horizon behavior in llms. *CoRR*, abs/2604.21748.

Buqiang Xu, Zirui Xue, Dianmou Chen, Chenyang Fu, Chiyu Wu, Caiying Huang, Chen Jiang, Jizhan Fang, Xinle Deng, Yijun Chen, Yunzhi Yao, Xuehai Wang, Jin Shang, Gong Yu, and Ningyu Zhang. 2026b. Tokenpilot: Cache-efficient context management for LLM agents. *CoRR*, abs/2606.17016.

Haiyang Xu, Qinghao Ye, Xuan Wu, Ming Yan, Yuan Miao, Jiabo Ye, Guohai Xu, Anwen Hu, Yaya Shi, Guangwei Xu, Chenliang Li, Qi Qian, Maofei Que, Ji Zhang, Xiao Zeng, and Fei Huang. 2023. Youkumplug: A 10 million large-scale chinese video-language dataset for pre-training and benchmarks. *CoRR*, abs/2306.04362.

Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. 2023. Clip-vip: Adapting pre-trained image-text model to video-language alignment. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

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, Zhongang Cai, Bo Li, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, and 3 others. 2025. 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.

Chongrui Ye, Yuxiang Liu, Yu Wang, Haofei Yu, Yining Zhao, Ge Liu, Julian McAuley, and Jiaxuan You. 2026. Auto-dreamer: Learning offline memory consolidation for language agents. *arXiv preprint arXiv:2605.20616*.

Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. 2023. Hitea: Hierarchical temporal-aware video-language pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 15359–15370. IEEE.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. 2025. Worldmm: Dynamic multimodal memory agent for long video reasoning. *CoRR*, abs/2512.02425.

Xinlei Yin, Xiulian Peng, Xiao Li, Zhiwei Xiong, and Yan Lu. 2026. Hierarchical long video understanding with audiovisual entity cohesion and agentic search. *CoRR*, abs/2601.13719.

Yufei Yin, Qianke Meng, Minghao Chen, Jiajun Ding, Zhenwei Shao, and Zhou Yu. 2025. Videoarm: Agentic reasoning over hierarchical memory for long-form video understanding. *CoRR*, abs/2512.12360.

Jeffrey M Zacks, Nicole K Speer, Khena M Swallow, Todd S Braver, and Jeremy R Reynolds. 2007. Event perception: a mind-brain perspective. *Psychological bulletin*, 133(2):273.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. 2025a. Videollama 3: Frontier multimodal foundation models for image and video understanding. *CoRR*, abs/2501.13106.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*, pages 543–553. Association for Computational Linguistics.

Kangning Zhang, Shuai Shao, Qingyao Li, Jianghao Lin, Lingyue Fu, Shijian Wang, Wenxiang Jiao, Yuan Lu, Weiwen Liu, Weinan Zhang, and Yong Yu. 2026a. Mmskills: Towards multimodal skills for general visual agents. *arXiv preprint arXiv:2605.13527*.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2025b. Long context transfer from language to vision. *Trans. Mach. Learn. Res.*, 2025.

Zeyu Zhang, Ziliang Guo, Yihang Sun, Xichong Zhang, Xixuan Hao, Zehao Lin, Yang Zhang, Xiaoyan Zhao, Tong Shen, Bo Tang, and 1 others. 2026b. Metis: Memory foundation model. *arXiv preprint arXiv:2607.26760*.

Appendix

A Dataset Details

This section provides additional details about the benchmark datasets used in the experiments. As shown in Table 7, we evaluate EM^{2}Mem on three long-video question answering benchmarks, covering both ultra-long egocentric daily-life videos and general long-form videos. EgoLifeQA and Ego-R1 Bench focus on first-person daily-life reasoning over hour-level video histories, while Video-MME (L) evaluates general long-video understanding across diverse visual domains.

Dataset	#Queries	Domain	Avg. Video Length
EgoLifeQA	500	Egocentric	44.3h
Ego-R1 Bench	300	Egocentric	44.3h
Video-MME (L)	900	General	0.69h

Table 7: Summary of benchmark datasets used in our experiments.

A.1 EgoLifeQA

EgoLifeQA (Yang et al., 2025) is a long-context egocentric video question answering benchmark built on week-long first-person daily-life recordings. It is designed to evaluate whether a model can serve as a personalized life assistant by retrieving and reasoning over sparse evidence distributed across long video histories. Following prior work, we evaluate on the A1_JAKE subject stream, which contains 44.3 hours of video and 500 multiple-choice questions.

The benchmark covers five types of memory-intensive queries. *EntityLog* questions require tracking objects, entities, and their states; *EventRecall* focuses on recalling specific past events and their temporal context; *HabitInsight* requires identifying recurring behaviors or long-term preferences; *RelationMap* evaluates understanding of social interactions and relationships; and *TaskMaster* asks about ongoing plans or pending tasks. This dataset is therefore well aligned with our goal of evaluating whether structured event memory can support object-centric, event-centric, and long-term semantic reasoning in ultra-long egocentric videos.

A.2 Ego-R1 Bench

Ego-R1 Bench (Tian et al., 2025) evaluates ultra-long egocentric video reasoning with 300 multiple-choice questions. It shares the daily-life video setting with EgoLifeQA, but emphasizes multi-step evidence gathering and tool-augmented reasoning over long video histories. To enable consistent analysis across egocentric benchmarks, we map its original query types to the five EgoLifeQA categories, as shown in Table 8.

Category	Ego-R1 Query Types
EntityLog	EntityLog, FoodLog, HealthLog, TechLog
EventRecall	EventRecall, Event Recollection, Event Memory
HabitInsight	HabitInsight, Behavior Habit(s)
RelationMap	RelationMap, Interpersonal Relationships
TaskMaster	TaskMaster, Future Plan(s)

Table 8: Mapping Ego-R1 Bench query types to the EgoLifeQA category taxonomy.

A.3 Video-MME

Video-MME (Fu et al., 2025) is a comprehensive benchmark for evaluating MLLMs on general video understanding, covering diverse domains, temporal durations, and multimodal inputs. Different from EgoLifeQA and Ego-R1 Bench, which focus on egocentric daily-life memory, Video-MME provides an open-domain evaluation setting for long-form video comprehension. In our experiments, we use only the long-video subset, denoted as Video-MME (L), which contains videos longer than 30 minutes and 900 multiple-choice questions. We adopt the benchmark's original task types.

A.4 License

We use only publicly available benchmarks released by their original authors. All datasets are used for academic evaluation under their corresponding licenses and terms of use. We do not redistribute original videos, annotations, or benchmark files, and users should obtain the datasets from the official sources and comply with the original licenses. For benchmarks with non-commercial or academic-use restrictions, such as Video-MME and EgoLife, our use is limited to research purposes; for permissively licensed datasets such as Ego-R1-Data, we follow their license terms.

B Implementation Details

This section provides additional details regarding the baseline implementations, the configuration of EM^{2}Mem, and the prompts employed in our experiments.

B.1 Baseline Setup

We compare EM^{2}Mem with a broad set of baselines covering four categories.

Base MLLMs. We include GPT-5 (OpenAI, 2026), Gemini 2.5 Pro (Team, 2025a), and Qwen3-VL-8B-Instruct (Team, 2025b). These models are evaluated as general-purpose multimodal LLMs without an external long-video memory module.

Long-video MLLMs. We compare with VideoChat-Flash (Li et al., 2025b), Time-R1 (Wang et al., 2025b), and Video-RTS (Wang et al., 2025c), which process sampled visual inputs within their respective context or frame-budget constraints. These baselines test whether stronger long-video input processing alone is sufficient for long-horizon QA.

RAG-based methods. We include LightRAG (Guo et al., 2025), HippoRAG (Gutierrez et al., 2024), and Video-RAG (Luo et al., 2024). LightRAG and HippoRAG retrieve from text-form video evidence such as captions or summaries, while Video-RAG retrieves relevant video clips before answer generation. These methods represent retrieve-then-generate pipelines without event-centric memory construction.

Memory-based long-video reasoning. We compare against EgoRAG (Yang et al., 2025), Ego-R1 (Tian et al., 2025), HippoMM (Lin et al., 2025), M3-Agent (Long et al., 2025), and WorldMM (Yeo et al., 2025). These methods maintain external memories or agentic retrieval procedures for long-video reasoning, making them the closest family of baselines to EM^{2}Mem.

Reported and reproduced results. For baselines other than WorldMM, we use the results reported in WorldMM (Yeo et al., 2025) on the overlapping benchmarks, including EgoLifeQA, Ego-R1 Bench, and Video-MME (L), to keep dataset splits, metrics, and evaluation protocols consistent. Since WorldMM is the strongest and most related memory-based baseline, we additionally reproduce it under the same evaluation setting as EM^{2}Mem and denote the reproduced version as WorldMM†.

WorldMM† reproduction. We follow the original WorldMM configuration wherever possible. To ensure a controlled comparison with EM^{2}Mem, WorldMM† uses the same backbone as our method when applicable: GPT-5-mini-2025-08-07 (OpenAI, 2026) is used during memory construction, including episodic and semantic memory construction and LLM-based triplet consolidation, while GPT-5-2025-08-07 is used as the retrieval agent and final answer generator. This corresponds to the WorldMM-GPT setting in the original paper. For episodic memory, we use temporal scales of 30 seconds, 3 minutes, 10 minutes, and 1 hour for EgoLifeQA and Ego-R1 Bench, and 10 seconds, 30 seconds, 3 minutes, and 10 minutes for Video-MME (L). For semantic memory, triplets with a similarity score greater than 0.6 are consolidated using an LLM, and the top-10 most relevant triplets are retrieved during inference. We keep the original visual memory design with both feature-based retrieval and timestamp-based frame access. The retrieval agent is allowed up to five iterations, and the remaining memory construction, retrieval, and response generation pipeline is kept unchanged.

B.2 EM^{2}Mem

Memory Construction Details. EM^{2}Mem is implemented as a training-free framework without updating model parameters. During memory construction, we use GPT-5-mini-2025-08-07 to convert event anchors into multimodal event records containing captions, transcripts, structured metadata, visual fields, timestamps, and provenance. We build temporal context views at 30 seconds, 3 minutes, 10 minutes, and 1 hour for EgoLifeQA and Ego-R1 Bench, and at 10 seconds, 30 seconds, 3 minutes, and 10 minutes for Video-MME (L). Episodic triplets are extracted from both local event records and temporal context views to construct the episodic graph. For semantic memory, we follow WorldMM (Yeo et al., 2025): triplets with similarity above 0.6 are consolidated with an LLM.

Inference Details. During inference, EM^{2}Mem retrieves evidence through event anchors with $K_E = 5$ episodic candidates and $K_S = 8$ semantic facts. Each candidate event e_i is scored using the local multimodal event record R_i, temporal context views C_i, episodic graph evidence G_i, and semantic facts S_i. We assign weight 1.00 to the local event record, use scale weights $0.65/0.45/0.30$ for 3-minute/10-minute/1-hour context views, and apply a one-hop graph expansion decay of 0.60. The LLM selector, implemented with GPT-5-2025-08-07, selects $K_{\text{sel}} = 5$ event anchors; for each selected anchor, we attach up to $K_V = 3$ keyframes

for final answer generation with GPT-5. Experiments are run on two NVIDIA A100 40GB GPUs for embedding-index storage and retrieval, with 8 parallel workers during inference evaluation.

For reproducibility, we provide a repository containing our implementation, prompts, configuration files, and evaluation scripts at https://github.com/zjunlp/LightMem.

B.3 Prompts

We report the prompt templates used in EM^{2}Mem for memory construction and inference. For memory construction, we design prompts for constructing Multimodal Event Records and Temporal Context Views. The text-side Multimodal Event Record prompt (Figures 6 and 7) guides the model to refine short event captions and extract structured fields such as salient objects, main actions, and conversation focus from captions and transcripts. The visual-side Multimodal Event Record prompt (Figures 8 and 9) instructs the model to annotate keyframes with visually grounded fields, including scenes, keyframe captions, and visual objects.

For Temporal Context Views, the text summary prompt (Figures 10 and 11) aggregates neighboring event records into coherent multi-scale textual summaries, while the visual summary prompt (Figures 12 and 13) summarizes visual fields across temporal windows. Together, these prompts enable event-centered memories at multiple granularities.

For inference, the LLM Selector prompt (Figures 14–18) selects the most relevant event anchors from retrieved event-centered candidates. The Answer Model prompt (Figure 19) generates the final answer based on the selected structured evidence and attached visual keyframes.

C Additional Description on Experiments

C.1 Offline Memory Construction Cost

Table 9 reports the token cost of offline memory construction. EM^{2}Mem follows an align-then-retrieve design that moves multimodal alignment and evidence organization from inference time to memory construction time. This introduces additional offline construction tokens, but organizes long-video evidence into compact event-centered memory units, enabling the inference stage to retrieve and reason over selected structured evidence instead of repeatedly aligning scattered modality-specific fragments. Since this cost is incurred once for each long-video memory, it can be amortized across repeated queries; as shown in Table 2, the resulting event-indexed memory leads to a more lightweight inference process with lower per-query cost and higher efficiency.

Method / Stage	Prompt	Completion	Total
WorldMM	23.26M	37.29M	60.55M
EM^{2}Mem			
Event Records	20.66M	12.30M	32.96M
Context Views	7.32M	2.58M	9.90M
Episodic Graph	9.07M	16.85M	25.92M
Semantic Memory	10.38M	10.08M	20.45M
EM^{2}Mem Total	47.43M	41.81M	89.23M

Table 9: Offline token cost of memory construction. Prompt and completion denote model input and output tokens. The construction cost is incurred once and amortized across repeated queries.

C.2 End-to-End Cost and Amortization Analysis

End-to-end cost and amortization. EM^{2}Mem intentionally shifts part of the computation from query-time reasoning to offline memory construction. This design is motivated by the observation that long-video memories are typically constructed once but queried multiple times. Table 10 reports the end-to-end wall-clock cost on EgoLifeQA, including both memory construction and inference. Compared with WorldMM, EM^{2}Mem increases the offline construction time from 88,708s to 99,022s, introducing an additional 10,314s construction overhead. However, this extra offline cost is substantially smaller than the inference-time saving: EM^{2}Mem reduces the inference wall-clock time from 229,502s to 6,138s. As a result, the overall end-to-end wall-clock time decreases from 318,210s to 105,160s, yielding a 3.03× speedup.

Metric	WorldMM	EM^{2}Mem	Δ / Gain
Wall-clock time			
Memory-Cons. (s)	88,708	99,022	+10,314
Inference (s)	229,502	6,138	-223,364
End-to-End (s)	318,210	**105,160**	**3.03**×
Token cost			
Memory-Cons	60.55M	89.23M	+28.68M
Inference	42.03M	15.27M	-26.76M
Total	102.58M	104.50M	+1.92M

Table 10: End-to-end cost on EgoLifeQA. "Memory-Cons." denotes offline memory construction, and "Inference" denotes QA inference. EM^{2}Mem incurs higher construction cost but substantially reduces inference cost, yielding a 3.03× end-to-end wall-clock speedup. The token-count break-even occurs after approximately 536 queries.

When does offline memory construction pay off. Let ΔT_{build} denote the additional construction time of EM^{2}Mem over WorldMM, and let ΔT_{infer} denote the per-query inference-time saving. The number of queries required to amortize the additional construction cost is

$$N_{\text{break-even}} = \frac{\Delta T_{\text{build}}}{\Delta T_{\text{infer}}}.$$

Using the measured average latency per query, EM^{2}Mem saves 360.79s per query, leading to a break-even point of approximately 29 queries. Using the practical evaluation throughput measured on EgoLifeQA, the break-even point is approximately 24 queries. This indicates that the proposed construction-time organization is beneficial when a processed long-video memory is reused for a moderate number of questions, such as benchmark evaluation, personal lifelog QA, surveillance-free archival search, or enterprise video repositories. In contrast, for one-off queries over videos that will never be revisited, the additional construction cost may not be fully amortized.

Token-cost amortization. We also analyze the same trade-off in terms of token usage. As shown in Table 10, EM^{2}Mem uses 28.68M more tokens than WorldMM during memory construction, but saves 26.76M tokens during inference on the 500-query EgoLifeQA evaluation set. Under an unweighted token-count calculation, this corresponds to a break-even point of approximately 536 queries. Therefore, EM^{2}Mem is immediately advantageous in wall-clock efficiency, while its token-cost advantage becomes clearer when the constructed memory is reused across a larger number of queries. Moreover, since EM^{2}Mem reduces output tokens more aggressively than input tokens during inference, the monetary break-even point can be lower when completion tokens are more expensive.

C.3 Detailed Ablation Study

Setup. Table 11 reports category-wise ablation results on EgoLifeQA. We compare the full EM^{2}Mem with four variants: *-TCV* removes temporal context views and only uses event-level evidence without coarser temporal aggregation; *-SM* removes semantic memory, so long-term facts and habits are not used as supporting evidence; *-EG* removes episodic graph evidence and disables graph-based event expansion; and *30s* restricts retrieval to 30-second event records only.

Type	Full	w/o TCV	w/o SM	w/o EG	30s-only
EntityLog	60.8	57.6	56.8	56.8	**61.6**
EventRecall	61.1	60.3	59.5	58.7	**65.1**
HabitInsight	**63.9**	60.7	62.3	60.7	60.7
RelationMap	**72.8**	60.8	65.6	68.8	64.8
TaskMaster	**74.6**	65.1	65.1	63.5	68.3
Overall	**66.0**	60.4	61.4	61.6	64.0

Table 11: Category-wise ablation on EgoLifeQA. Full denotes EM^{2}Mem. w/o TCV, w/o SM, and w/o EG remove temporal context views, semantic memory, and episodic graph, respectively. 30s uses only 30-second event records. Accuracy is reported in %.

Analysis. The full model achieves the best overall accuracy, outperforming all ablated variants by 2.0–5.6 points. The gains are most evident on HabitInsight, RelationMap, and TaskMaster, where questions require recurring patterns, interpersonal relations, or task-level continuity beyond a single short event. Removing temporal context views causes the largest overall drop, showing the importance of coarse context for connecting sparse evidence across long videos. Removing semantic memory and episodic graph also degrades performance, indicating that long-term semantic support and event-level relational links are both useful for evidence-driven reasoning. The 30s-only variant performs best on EntityLog and EventRecall, suggesting that fine-grained event records are sufficient for local object-centric or direct recall questions, but its weaker results on long-range categories confirm the need for temporal context, episodic graph structure, and semantic memory.

C.4 Analysis of Construction-Time Alignment

WorldMM (Yeo et al., 2025), our strongest and most related memory-based baseline, also employs multi-scale, semantic, visual, and graph-based memories. To isolate the contribution of construction-time evidence alignment from these shared components, we compare two controlled variants on all 500 EgoLifeQA questions using the same 6,057 underlying 30-second records. Both use identical caption, transcript, visual, and structured evidence, retrieval and answer models, prompts, causal filtering, and temporal budgets, while disabling multi-scale context, graph expansion, LLM selection, and query planning. *Four-channel RRF* indexes the four evidence channels independently and combines their rankings with reciprocal-rank fusion, whereas *Aligned-30s* binds the same fields into one provenance-linked record per temporal anchor before indexing.

Memory organization	$R@1$	$R@5$	$C@150$	$C@300$	$C@600$	QA
4-channel RRF	8.9	24.2	24.3	31.2	41.1	60.0
Aligned-30s	**14.0**	**27.6**	**27.9**	**36.7**	**44.6**	**63.0**
Gain	+5.1	+3.4	+3.6	+5.4	+3.5	+3.0

Table 12: Controlled comparison of independent four-channel retrieval and construction-time aligned retrieval on all 500 EgoLifeQA questions. $R@k$ denotes top-k target-time recall, $C@B$ denotes macro target-time coverage under a B-second retrieved-video budget, and QA denotes final-answer accuracy (%).

As shown in Table 12, construction-time alignment consistently improves evidence localization and coverage. Aligned-30s improves $R@1$ and $R@5$ by 5.1 and 3.4 points, respectively, and increases coverage across all temporal budgets. At the primary 300-second budget, coverage improves by 5.4 points (95% CI: [+2.6, +8.3]; McNemar exact $p = 0.0003$). Final QA accuracy also increases from 60.0% to 63.0%, although this difference is not statistically significant (95% CI: [-0.8, +6.8]; $p = 0.1505$). These results indicate that EM^{2}Mem's advantage is not solely attributable to auxiliary memory components shared with prior systems: binding heterogeneous evidence into a common retrieval unit before indexing improves evidence retrieval under matched conditions.

C.5 Category-wise Evidence Unification Analysis

This section extends Section 4.4 with category-wise results on EgoLifeQA. Table 13 shows that EM^{2}Mem achieves the best overall accuracy, but the benefit is not uniform across question types. The largest gains appear on *RelationMap* and *TaskMaster*, where the model must connect people, actions, tasks, and temporally distant evidence across multiple events. This suggests that the main advantage of construction-time unification is not simply adding visual information, but turning heterogeneous cues into event-centered memory units that can be retrieved and composed as coherent evidence chains. In contrast, *EventRecall* is relatively insensitive to the choice of evidence interface, indicating that direct recall questions can often be answered from localized evidence without requiring strong cross-event integration.

The results also reveal the boundary of structured memory. Structured event records perform strongly on *EntityLog*, suggesting that typed fields such as objects, actions, and scenes provide an effective interface for local entity-centric reasoning. However, raw visual evidence performs best on *HabitInsight*, implying that some recurring perceptual cues may be weakened when visual information is compressed into textual or structured fields. This observation supports our final design: EM^{2}Mem uses structured event records as the primary retrieval interface for efficient and traceable reasoning, while retaining raw keyframes only for targeted visual verification after relevant event-centric multimodal memory selection. Overall, the category-wise trends refine the main conclusion in Section 4.4: effective multimodal memory depends on both *how* evidence is represented—as structured event-centered memory—and *when* it is unified—before retrieval rather than during inference.

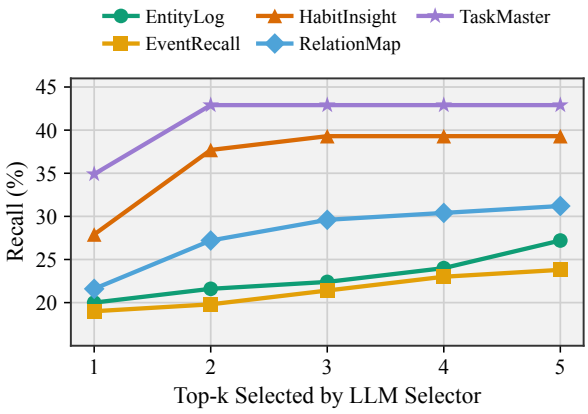


Figure 4: Event-level recall under different selector budgets.

C.6 Selector budget

We evaluate whether EM^{2}Mem retrieves the annotated temporal evidence more effectively than WorldMM. We use strict 30-second event-level recall, where a prediction is counted as correct only when the retrieved event anchor matches the annotated evidence segment. Table 6 and Figure 4 compares WorldMM's cumulative recall after 1 and 5 iterative retrieval rounds with EM^{2}Mem's single-pass Top-k event-anchor retrieval. EM^{2}Mem achieves 30.8% Top-5 recall, outperforming WorldMM 5R by 7.0 points. EM^{2}Mem Top-1 recall also reaches 23.0%, close to WorldMM 5R at 23.8%. The gains are larger on HabitInsight and TaskMaster, suggesting that event-indexed memory is helpful when evidence depends on long-term patterns or task states. EventRecall is the only category where EM^{2}Mem is slightly lower, indicating that localized event-recall

Question Type	#Questions	Cap.-Late	Cap.-Early	Vis.-Late	Vis.-Early	SER-Late	EM^{2}Mem
EntityLog	54	59.3	63.0	57.4	64.8	**68.5**	64.8
EventRecall	45	68.9	68.9	66.7	68.9	68.9	68.9
HabitInsight	40	65.0	65.0	**72.5**	67.5	62.5	65.0
RelationMap	71	66.2	64.8	70.4	73.2	69.0	**74.6**
TaskMaster	40	67.5	67.5	67.5	67.5	70.0	**77.5**
Overall	250	65.2	65.6	66.8	68.0	68.0	**71.2**

Table 13: Analysis of evidence interface and alignment timing on EgoLifeQA. Cap., Vis., and SER denote flat caption evidence, raw visual evidence, and structured event records, respectively. Late denotes retrieval-time alignment, while Early denotes construction-time alignment. EM^{2}Mem denotes our event-centric unified multimodal memory framework. All numbers denote accuracy (%).

questions may still benefit from iterative caption-level retrieval. Overall, these results suggest that EM^{2}Mem improves evidence localization for structured and long-range reasoning while preserving competitive localized recall.

C.7 Anchor Granularity and Boundary Sensitivity

The base anchor in EM^{2}Mem is a fixed temporal indexing address rather than a learned semantic boundary. We evaluate its granularity on all 500 EgoLifeQA questions by comparing fixed 30-, 60-, and 120-second windows with two variable-length proxies that merge consecutive 30-second records while the generated scene label or speaker set remains unchanged. All variants use the same evidence and retrieval settings. Since longer windows cover more video per retrieved item, we report macro target-time coverage under equal retrieved-video budgets of 150, 300, and 600 seconds.

Anchor construction	C@150	C@300	C@600
Fixed 30s	**27.9**	**36.7**	**44.6**
Fixed 60s	20.2	30.3	39.4
Fixed 120s	16.5	23.9	34.4
Scene-boundary proxy	17.6	26.0	36.2
Speaker-boundary proxy	26.8	33.3	42.3

Table 14: Anchor sensitivity on all 500 EgoLifeQA questions. $C@B$ denotes macro target-time coverage (%) under a B-second retrieved-video budget.

As shown in Table 14, fixed 30-second anchors achieve the highest coverage at all budgets; at 300 seconds, they outperform the alternatives by 3.4–12.8 points. We also tested identity-based adaptive boundaries using YOLOv8 with BoT-SORT on one full day of EgoLife video. Despite only six protagonists, tracking produced 6,130 person IDs, with frequent fragmentation under egocentric motion, occlusion, and re-entry. These results suggest that 30-second anchors offer a robust localization–budget trade-off on EgoLifeQA, without implying a universal semantic event boundary.

C.8 Case Study

Figure 5 presents a qualitative case study on an EgoLifeQA TaskMaster question: "Who plans to grow flowers?" The answer requires identifying plan ownership in a multi-person discussion where flowers, vases, craft ideas, and nearby mentions of Tasha appear together. WorldMM retrieves several flower-related snippets, but they remain loosely connected textual fragments without an explicit person–plan binding, leading to the wrong prediction Tasha. In contrast, EM^{2}Mem aligns multimodal evidence during memory construction and stores it as an event-centric memory cell, where Katrina is explicitly linked to the flower-based plan, associated objects, temporal context, and supporting visual evidence.

Table 15 further traces how this evidence propagates through the pipeline. The local record and episodic graph preserve Katrina-specific person–plan relations with event provenance, while retrieved Tasha-related semantic facts provide plausible but non-decisive distractors. After causal ranking and bounded selection, the retained Katrina events contain both the 7-day cultivation context and the explicit statement "We can plant it too," whereas Alice-related evidence concerns bulbs purchased as prizes. The grounded readout therefore resolves the ambiguous flower mentions to Katrina and yields the correct answer. This example illustrates how event-centric alignment, provenance-linked graph structure, and bounded retrieval jointly turn heterogeneous evidence into a query-ready evidence unit.

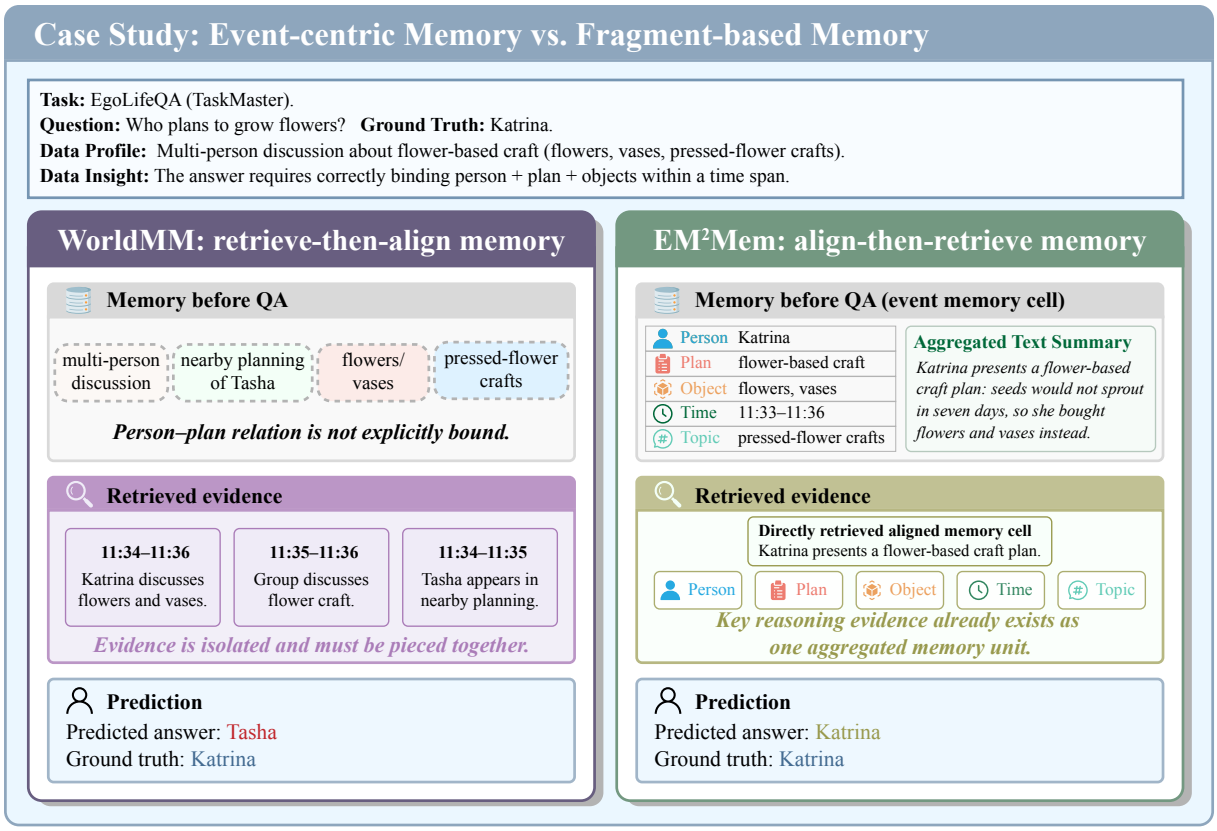


Figure 5: Case study of TaskMaster in EgoLifeQA.

D Extended Related Work

Long Video Understanding. Recent video-language models have achieved substantial progress in processing extended visual sequences. Early video-language pretraining methods learn transferable visual-semantic representations through joint video-text modeling, multimodal alignment, and masked spatiotemporal modeling (Sun et al., 2019; Li et al., 2022a; Tong et al., 2022; Xue et al., 2023; Wang et al., 2024b). Long-form and temporal-aware pretraining further improve video-language modeling over extended temporal contexts (Sun et al., 2022; Ye et al., 2023; Jin et al., 2022; Xu et al., 2023; Han et al., 2023). Building on these foundations, video large language models connect visual encoders with LLMs and improve instruction-following through multimodal instruction tuning and unified image-video representations (Zhang et al., 2023; Maaz et al., 2024a; Lin et al., 2024; Jin et al., 2024; Li et al., 2025a). Recent systems further strengthen caption quality, audio-spatiotemporal modeling, and image-video encoder integration (Chen et al., 2024; Cheng et al., 2024; Maaz et al., 2024b; Zhang et al., 2025a; Wang et al., 2024a). More recent work extends video LLMs to longer contexts by compressing visual tokens, maintaining sparse memory, or adapting visual inputs to long-context LLMs (Song et al., 2024; Li et al., 2024; Zhang et al., 2025b; Shen et al., 2025; Li et al., 2025b). Other studies improve temporal localization and reasoning through time-aware modeling, temporal grounding, or test-time reasoning (Ren et al., 2024; Wang et al., 2025b,c,a). Nevertheless, as video duration increases from minutes to hours or days, task-relevant evidence often becomes sparse, temporally distant, and difficult to localize, motivating explicit mechanisms for organizing and retrieving video-specific information.

Multimodal Memory for Large Language Models. To enable scalable reasoning over long videos, recent studies have explored explicit memory and retrieval mechanisms. Retrieval-augmented methods build external memories from captions, transcripts, OCR, keyframes, clip embeddings, or graph-structured indices, and retrieve query-relevant evidence before answer generation (Guo et al., 2025; Gutierrez et al., 2024; Luo et al., 2024; Jeong et al., 2025; Kahatapitiya et al., 2025). Other methods organize videos into document-like, hierarchical, or adaptive structures for efficient

Stage or object	Actual stored or retrieved evidence	Role in the trace
Local record	At `DAY1_11333000_11340000`, Katrina says that she bought flowers and vases, that the flowers will bloom in 1–2 days, and that this fits the 7-day cycle. A linked keyframe independently shows pink flowers in a vase.	Binds the speaker, intended activity, time span, and visual observation to one source interval.
Episodic graph	The node `event::DAY1_11333000_11340000` `involves` Katrina, is `about` buying flowers and vases, and `occurs_in` the dining area. Extracted triplets include (Katrina, `say_about`, bought flowers and vases) and (Katrina, `say_about`, fits 7-day cycle).	Preserves the person–plan association. Every relation retains its event ID, document ID, and 30s scale for source tracing.
Semantic distractors	Retrieved facts also state that Tasha says seed paper can be planted (support count 5) and that sunflower seeds are easier to sprout (support count 1).	These facts overlap with *grow flowers*, but do not identify the owner of the queried plan; they enter only through their provenance-linked anchors.
Ranking and selection	With a causal cutoff at 11:57:36, the selector retains five anchors from the 12-anchor pool: 11:35:30 (1.0000), 11:47:30 (0.9288), 11:47:01 (0.8839), 11:48:00 (0.4272), and 11:34:30 (0.3954).	The scores are produced by coarse event ranking; bounded selection compares the complete event packets without searching the full memory.
Grounded readout	The 3-minute view around 11:35 contains Katrina's 7-day cultivation plan, while the 11:48 event adds her statement, "We can plant it too." Later Alice events concern bulbs purchased as prizes rather than her own growing plan.	The assembled evidence resolves plan ownership and yields the correct answer, **D: Katrina**.

Table 15: Concrete memory objects and query-time trace for the question "Who plans to grow flowers?" Scores in the ranking row are normalized coarse anchor scores.

long-video reasoning (Ma et al., 2025; Wang et al., 2025d; Chen et al., 2025; Xu et al., 2026a; Yin et al., 2026). In egocentric long-video understanding, EgoRAG constructs multi-level caption and summary memories for coarse-to-fine retrieval (Yang et al., 2025; Deng et al., 2026a,b), while HippoMM segments audiovisual streams into episodes and consolidates them into semantic summaries for hierarchical recall (Lin et al., 2025). Beyond fixed retrieval pipelines, agentic memory systems incorporate memory access into multi-step reasoning for temporal localization, multimodal inspection, and iterative evidence gathering (Fan et al., 2024; Long et al., 2025; Tian et al., 2025; Yeo et al., 2025; Yin et al., 2025; Li et al., 2026a). Collectively, this line of work (Wen et al., 2026; Liang et al., 2026; Lian et al., 2026; Fang et al., 2025; Xu et al., 2026b; Chen et al., 2026b; Cao et al., 2026) motivates the development of unified memory systems capable of supporting fine-grained event grounding, flexible temporal abstraction, and long-term semantic reasoning (Zhang et al., 2026a; Qiao et al., 2023; Nie et al., 2026; Ye et al., 2026; Chen et al., 2026a; Zhang et al., 2026b; Li et al., 2026b).

Prompts for Multimodal Event Record
(Texual Evidence)

[SYSTEM]:
Role and Objective
You are an expert video captioner and evidence annotator.
Your task is to transform caption and transcript segments for a short video clip into:
1. a concise, item-focused fine-grained caption that clearly describes key actions, objects, and interactions;
2. structured evidence fields that summarize salient objects, main actions, and conversation focus.
Input Data Format
You will receive a JSON object with:
- `wearer_name`: optional camera wearer name for egocentric videos
- `original_worldmm_caption`: the caption already generated by WorldMM
- `segments`: a JSON array where each element has:
  - `start`: start timestamp
  - `end`: end timestamp
  - `text`: caption or transcript text
  - `type`: either `caption` or `transcript`
Some clips may contain only captions, only transcripts, or both.
Caption Handling
- Use caption segments to anchor the visible actions, objects, tools, and interactions.
- Refine or merge redundant caption content for clarity.
Transcript Handling
- Preserve other speakers' names (e.g., `Shure:`).
- If the transcript speaker is the camera wearer, you may remove the wearer's name and render the speech naturally.
- If no transcript is provided, do not invent dialogue or conversation topics.
Guidelines
1. Output every field in natural English only.
2. If the source is egocentric or clearly first-person, keep a first-person perspective. Otherwise, preserve the perspective already implied by `original_worldmm_caption` and do not force first-person wording.
3. Highlight key objects, tools, or interactions.

Figure 6: Prompt for constructing text-derived fields in Multimodal Event Records, part 1.

Prompts for Multimodal Event Record
(Texual Evidence)

4. Keep captions concise, direct, and focused on items or actions.
5. Merge overlapping or redundant content.
6. Ensure the final caption corresponds to the original sequence.
7. Favor clarity and succinctness over extensive narration.
8. `salient_objects` should contain concrete objects, tools, or notable items.
9. `main_actions` should be short verb phrases.
10. `conversation_focus` should summarize the main topics or requests in speech, not full transcripts. If there is no meaningful transcript evidence, return an empty list.
11. Do not invent events, objects, or speech that are not supported by the input.
12. Return valid JSON only.
Output JSON Format
Return exactly:
```
{
  "fine_caption": "string",
  "salient_objects": ["..."],
  "main_actions": ["..."],
  "conversation_focus": ["..."]
}
```
[USER]:
```
{
  "wearer_name": "{wearer_name}",
  "original_worldmm_caption": "{original_worldmm_caption}",
  "segments": [
    {
      "start": "{start_timestamp}",
      "end": "{end_timestamp}",
      "text": "{caption_or_transcript_text}",
      "type": "caption | transcript"
    }
  ]
}
```

Figure 7: Prompt for constructing text-derived fields in Multimodal Event Records, part 2.

Prompts for Multimodal Event Record (Visual Evidence)

```
[ SYSTEM ]:
# Role and Objective
You are an expert visual annotator.
You will be given:
- a few keyframes from a short video segment;
- a short text context for grounding only.
Your job is to describe what is visually present in the keyframes, not to paraphrase the text context.
# Instructions
1. Focus on what is actually visible in the images.
2. Do not copy the text context unless it is visually supported.
3. Use concise English descriptions only.
4. `scene` should be a coarse environment label such as:
   kitchen, dining area, living room, bedroom, bathroom, hallway, office/workspace,
   outdoors, store/shop, restaurant/cafe, car/interior, other
5. `keyframe_caption` should describe the visible scene in a natural style that matches the clip perspective.
6. `visual_objects` should contain only clearly visible objects/items.
7. Return valid JSON only.
# Output JSON Format
Return exactly:
{
  "scene": "string",
  "keyframe_caption": "string",
  "visual_objects": ["..."]
}
```

Figure 8: Prompt for constructing visual fields in Multimodal Event Records, part 1.

Prompts for Multimodal Event Record (Texual Evidence)

```
[ USER ]:
{
  "text_context": {
    "worldmm_caption": "{fine_caption_or_original_worldmm_caption}",
    "caption_text": "{caption_text}",
    "transcript_text": "{transcript_text}"
  },
  "instruction": "Use the text only as weak context. Describe only what is visually supported by the images. Do not
paraphrase the text unless visible."
}

[ USER IMAGES ]:
{keyframe_1}
{keyframe_2}
...
{keyframe_n}
```

Figure 9: Prompt for constructing visual fields in Multimodal Event Records, part 2.

Prompts for Temporal Context Views
(Texual Summary)

[SYSTEM]:
As an Event Summary Documentation Specialist, your role is to systematically structure and summarize event information, ensuring that all key actions of major characters are captured while maintaining clear event logic and completeness. Your focus is on concise and factual summarization rather than detailed transcription.
Specific Requirements
1. Structure the Events Clearly
- Merge related lower-level events into a coherent 10-minute event summary.
- Arrange the content chronologically.
- Group information by ongoing task, discussion, or theme when appropriate.
2. Retain Key Information
- Preserve the main actions and decisions of "I".
- Preserve important interactions with other participants.
- Preserve recurring objects, discussion topics, and speakers only when they matter to the main event flow.
3. Use Aggregated Text Metadata as Supporting Constraints
- You will receive:
  - ordered_event_units: chronological child summaries
  - aggregated_text_metadata: filtered canonicalized recurring actions, objects, topics, speakers
- Use ordered_event_units as the main chronological backbone.
- Use aggregated_text_metadata to preserve important recurring details.
- Do NOT turn metadata lists into direct prose unless clearly necessary.
- You may also receive critical_speech lines extracted from transcript evidence.
- If critical_speech contains an explicit proposal, plan, commitment, instruction, source-transfer relation, or exact named item, preserve that information faithfully.
- Preserve exact nouns when they are important to the event meaning, instead of replacing them with broader categories.
- If a lower-level unit explicitly states where an object came from / was placed / was before, keep that source relation in the summary.
- If a lower-level unit explicitly states a proposal or intention, keep the proposer/agent and the proposed content together.
4. Concise Expression, Remove Redundancies
- Focus on major developments, task flow, and conclusions.
- Remove repeated details and overly local observations.
- Summarize into event stages instead of listing many individual actions.
- Do NOT include micro gestures, local posture changes, or low-level repetitive motions.

Figure 10: Prompt for constructing text summaries in Temporal Context Views, part 1.

Prompts for Temporal Context Views
(Texual Summary)

5. Strictly Adhere to Facts
- Do not speculate.
- Keep the summary factual and chronological.
Output Format
Return valid JSON only:
```
{
  "summary_text": "string"
}
```
[USER]:
```
{
  "ordered_event_units": [
    {
      "time": "{date}_{start_time}_{end_time}",
      "event": "{child_summary_text}",
      "critical_speech": [
        "{critical_speech_line_1}",
        "{critical_speech_line_2}"
      ]
    }
  ],
  "aggregated_text_metadata": {
    "action_threads": ["{filtered_canonicalized_recurring_actions}"],
    "object_threads": ["{filtered_canonicalized_recurring_objects}"],
    "topic_threads": ["{filtered_canonicalized_recurring_topics}"],
    "speaker_stats": ["{speaker_statistics}"],
    "protected_mentions": ["{protected_named_entity_or_exact_mention}"]
  }
}
```

Figure 11: Prompt for constructing text summaries in Temporal Context Views, part 2.

Prompts for Temporal Context Views
(Visual Summary)

[SYSTEM]:
You are a Visual Event Summary Documentation Specialist for egocentric videos.
Your job is to summarize the visual content of a medium time period based only on visual evidence.
Specific Requirements
1. Structure the Visual Content Clearly
- Use ordered visual descriptions as the main chronological backbone.
- Organize the visual summary around dominant scene context, recurring visible objects, and major visual transitions.
2. Retain Key Visual Information
- Preserve dominant scene context.
- Preserve stable and repeated visible objects.
- Preserve major visual changes only if clearly supported.
3. Use Aggregated Visual Metadata as Supporting Constraints
- You will receive:
  - ordered_visual_units: chronological child visual descriptions
  - aggregated_visual_metadata: filtered canonicalized scene statistics and visual objects
- Use ordered_visual_units as the primary visual backbone.
- Use aggregated_visual_metadata to preserve recurring visual context.
- Do NOT turn metadata lists into direct prose.
4. Concise Expression, Remove Redundancies
- Focus on stable visual context rather than local transient details.
- Avoid repeating similar visual observations.
- Do NOT mention low-information background objects unless they are central to the visual scene.
5. Strictly Adhere to Facts
- Only describe what is visually supported.
- Do not speculate.
Output Format
Return valid JSON only:
```
{
  "visual_summary": "string"
}
```

Figure 12: Prompt for constructing visual summaries in Temporal Context Views, part 1.

Prompts for Temporal Context Views
(Visual Summary)

[USER]:
```
{
  "ordered_visual_units": [
    {
      "time": "{date}_{start_time}_{end_time}",
      "visual_evidence": "{child_visual_summary_or_visual_backbone_text}"
    }
  ],
  "aggregated_visual_metadata": {
    "scene_summary": {
      "dominant_scene": "{dominant_scene}",
      "scene_distribution": {
        "{scene_label}": "{probability_or_ratio}"
      }
    },
    "visual_object_threads": [
      {
        "object": "{filtered_canonical_visual_object}",
        "canonical_label": "{filtered_canonical_visual_object}",
        "aliases": ["{raw_visual_object_alias}"],
        "support_units": "{support_count}",
        "first_seen": "{start_time}",
        "last_seen": "{end_time}",
        "child_ids": ["{child_doc_id}"],
        "child_coverage": "{coverage}"
      }
    ]
  }
}
```

Figure 13: Prompt for constructing visual summaries in Temporal Context Views, part 2.

Prompts for LLM Selector

[SYSTEM]:
You are selecting event packets for a long-video QA system.
Your job is NOT to choose events that are merely topically related. Your job is to choose events whose evidence matches the exact predicate asked by the question.
You must do two things:
Step 1: infer the question family from the question.
Step 2: choose a small, complementary set of event packets that best supports the answer.
Use one of these question families:
1) action-owner
2) source-trace
3) participant-membership
4) plan-intention-decision
5) temporal-recall
6) habit-preference
7) attribute-content-purpose
Core principle:
- Prefer explicit evidence over weak implication.
- Prefer predicate-aligned evidence over broad contextual relevance.
- Do not over-select near-duplicate local events.
- Always return valid candidate indices and/or valid doc_ids from the provided list only.
[action-owner]
Question intent: identify who performed an action, who assisted, or who acted first.
Strong evidence:
- explicit actor + explicit queried action
- explicit cooperation, transfer, or assistance evidence when the question is about helping
- earliest valid explicit action when the question is about who acted first
Weak evidence:
- nearby presence
- interaction with related objects without the queried action
- later result scenes without explicit action evidence
Do NOT:
- treat holding, using, or interacting with an object as sufficient evidence of prior location
- answer a previous-location question using only current-scene context
- omit a source-establishing event if one exists

Figure 14: Prompt for the LLM Selector, part 1.

Prompts for LLM Selector

[participant-membership]
Question intent: identify who joined, who helped, who was part of the activity, or who was absent.
Strong evidence:
- explicit participation in the shared activity
- explicit join/help/presence evidence in the relevant action chain
- contrastive evidence for absence or mismatch across time
Weak evidence:
- later co-presence in the same room
- nearby observer or bystander context
Do NOT:
- infer participation only from later appearance
- confuse bystanders with core participants
[plan-intention-decision]
Question intent: identify a plan, intention, decision, next step, proposal, or commitment.
Strong evidence:
- explicit plan, intention, decision, proposal, assignment, or commitment
- agent-specific future commitment
- final-decision evidence
Weak evidence:
- related discussion
- explanation, recommendation, or evaluation
- general topic proximity
- observation statements without commitment
- offer or suggestion unless it clearly implies the agent's own intended action
Do NOT:
- infer intention from discussion alone
- infer a personal plan from explanation or recommendation alone
- confuse proposal, observation, ownership, or topic relevance with intention

Figure 15: Prompt for the LLM Selector, part 2.

Prompts for LLM Selector

[temporal-recall]
Question intent: identify the last time, first time, previous occurrence, or temporally constrained event.
Strong evidence:
- event whose timestamp best satisfies the temporal constraint
- closest valid earlier or later occurrence that truly matches the queried event or topic
Weak evidence:
- semantically similar event at the wrong time
- salient but temporally invalid event
Do NOT:
- ignore first/last/before/after constraints
- choose a more relevant-looking event if its time is wrong
[habit-preference]
Question intent: identify a repeated behavior, usual pattern, stable preference, or dislike.
Strong evidence:
- repeated evidence across multiple events
- explicit preference statements
- aggregate frequency patterns
Weak evidence:
- one-off action
- isolated or accidental occurrence
Do NOT:
- infer a habit from only one weak event if stronger repeated evidence exists
- confuse temporary behavior with stable preference

Figure 16: Prompt for the LLM Selector, part 3.

Prompts for LLM Selector

[attribute-content-purpose]
Question intent: identify ownership, contents, identity, purpose, attribute, or category.
Strong evidence:
- direct statement of ownership, contents, identity, purpose, or queried attribute
- explicit visual or textual grounding of the queried property
Weak evidence:
- nearby action context
- related discussion without direct attribute grounding
Do NOT:
- replace a direct attribute question with surrounding activity
- infer ownership, content, purpose, or identity from loose association alone
Global anti-error rules:
- Do not infer agent ownership from scene participation alone.
- Do not infer intention from topic discussion alone.
- Do not infer source from current location alone.
- Do not infer habits from a single weak event if stronger repeated evidence exists.
- Do not infer attributes from nearby actions when direct grounding exists.
- When direct evidence and broad contextual evidence conflict, prefer direct evidence.
- Use role scores as hints, not hard constraints.
- Prefer a smaller set of directly relevant events over a larger set of vaguely related events.
- If a question has a critical constraint (actor, source, time, intention, ownership, identity), at least one selected event should directly ground that constraint.

Figure 17: Prompt for the LLM Selector, part 4.

Prompts for LLM Selector

[USER]:
{query_with_time}
Candidate Event Packets:
{selector_candidates}
Select the best {final_top_k} candidates.
Selection goals:
- Choose complementary evidence, not repetitive evidence.
- Retain at least one event that directly grounds the core predicate of the question.
- If the question requires prior-state or source evidence, retain the event that directly establishes that prior state, even if it is earlier and less salient.
- If the question requires intention or decision evidence, retain explicit commitment or decision evidence rather than topic-related discussion.
- If the question requires identifying an actor, retain explicit actor evidence.
- If the question requires temporal comparison, enforce the temporal constraint strictly.
- If the question requires a stable habit or preference, prefer repeated or aggregate evidence over one-off evidence.
- If the question requires ownership, contents, identity, purpose, or attribute, prefer direct grounding over surrounding context.
Output requirements:
- Infer the correct question_family first.
- Then select the best candidates.
- The reason must explain why the selected events satisfy the core predicate better than merely related events.
Return ONLY JSON in this format:
{
  "question_family": "...",
  "selected_indices": [..],
  "selected_doc_ids": [..],
  "reason": "..."
}

Figure 18: Prompt for the LLM Selector, part 5.

Prompts for LLM Answer

[SYSTEM]:
You are an AI assistant that answers questions about egocentric video experiences using retrieved memory context. Your task is to answer multiple choice questions based on this accumulated context. Always choose the most relevant answer from the given choices based on the evidence provided.
Guidelines
- Analyze all provided context carefully.
- Choose the answer that best matches the evidence.
- If evidence is unclear, make the most reasonable inference.
Output Format
Provide your answer as a single letter (A, B, C, or D) based on the evidence.
[USER]:
Query: {query}
Query Time: {query_time, optional}
Important: Interpret all relative temporal expressions (e.g. "before", "after", "earlier", "later", "recently", "a few hours ago", "first", "last") relative to Query Time.
Choices: {choices, optional}
Context:
Selector summary:
Chosen event anchors: {top_doc_ids}
Selector reason: {selector_reason}
The selected event anchors were chosen because they form the strongest evidence chain for this question.
Use these selected events as the primary basis for answering.
Do not override a clearly supported conclusion from the selected evidence with a weaker alternative.
{retrieved_event_packets}
{visual_evidence_images, optional}
{multiple_choice_grounding_instructions, optional}

Figure 19: Prompt for the final Answer Model.