EMMA: Event-Driven Multimodal Memory System with Multi-Agent Retrieval

Ruochen Zhao[1]*, Jiazheng Kang[1]*, Gezi Li[2], Dongying Lin[2], Ting Bai[1]†

[1]Beijing University of Posts and Telecommunications
[2]Huawei Technologies Co., Ltd.
{kjz,zhaorc,baiting}@bupt.edu.cn
{ligezi,lindongying}@huawei.com

Abstract

Long video understanding requires reasoning over hours of content that far exceed the context window of multimodal large language models (MLLMs). Existing approaches compress video into fixed-length clips, yet events are fragmented and unrelated content is merged, yielding noisy retrieval. To overcome this challenge, we propose **EMMA**, an **E**vent-driven **M**ultimodal **M**emory system with Multi-**A**gent retrieval, which organizes memory around semantic events in three stages: (1) Event segmentation: a dual-stream detector adaptively locates event boundaries by fusing visual semantic drift with spoken transitions, rather than fixed windows. (2) Memory construction: built on the detected events, a three-level hierarchy of raw video, clip-level memory, and event-level memory stores content at progressively higher semantic abstraction. (3) Memory retrieval: a multi-agent framework collaboratively queries this hierarchy through query decomposition, tiered search, evidence validation, and answer generation, narrowing evidence from coarse events to precise clips. Extensive experiments on three challenging benchmarks demonstrate substantial improvements, showing the effectiveness of our event-aligned memory system. Code, prompt templates, and reproduction scripts are available at https://github.com/BAI-LAB/EMMA.

1 Introduction

Long video understanding requires reasoning over hours of visual content that forms a coherent storyline. Each frame is embedded in a continuous narrative structure, yet multimodal large language models (MLLMs) (Liu et al., 2023; Bai et al., 2025) cannot ingest such lengthy input in a single pass. Existing memory-based systems (Fan et al., 2024; Zuo et al., 2026; Long et al., 2025) compress video into retrievable units, but they share a fundamental design choice: unit boundaries are decided by predefined temporal parameters rather than by the video's event structure. This content-agnostic segmentation fractures the intrinsic storyline, causing two systematic retrieval failures. First, *evidence fragmentation*: a coherent event is split across adjacent units, so no single retrieval can reassemble the complete evidence. Second, *evidence dilution*: an event shorter than the preset window is merged with unrelated material, degrading semantic specificity.

Aligning memory units with the video's intrinsic event structure is a natural remedy, grounded in Event Segmentation Theory (EST) (Zacks et al., 2007): the human brain partitions ongoing experience into discrete, internally consistent event units to enable precise and efficient recall. Translating EST into a computational framework imposes three desiderata. *(D1) Semantic alignment*: boundaries must respond to content drift, not elapsed time. *(D2) Cross-modal robustness*: a boundary may appear visually or verbally alone, so a single modality is structurally insufficient. *(D3) Multi-granularity retrieval*: the QA layer must adaptively choose evidence granularity, from a short utterance to a whole event trajectory.

We instantiate these desiderata in **EMMA**, an **E**vent-driven **M**ultimodal **M**emory system with **M**ulti-**A**gent retrieval, through three coordinated components: **(1) Event segmentation** (*D1-D2*): a dual-stream detector fuses visual semantic drift with spoken chapter transitions to locate event boundaries adaptively, replacing fixed temporal windows. **(2) Memory construction**(*D3*): atop the detected events, a three-level hierarchy of raw video, clip-level memory (CLM), and event-level memory (ELM) stores content at progressively higher semantic abstraction. **(3) Memory retrieval** (*D3*): a multi-agent framework collaboratively queries this hierarchy through query decomposi-

*Equal contribution.
†Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 26549–26561
October 24-29, 2026 ©2026 Association for Computational Linguistics

tion, tiered search, evidence validation, and answer generation, narrowing evidence from coarse events to precise clips. Extensive experiments on three challenging long-video benchmarks demonstrate that **EMMA** achieves substantial improvements over strong baselines, showing the effectiveness of event-aligned memory organization. The contributions of our work are:

- We diagnose evidence fragmentation and evidence dilution caused by content-agnostic boundaries in existing memory-based video agents, and propose dual-stream event segmentation that adaptively locates event boundaries by fusing visual semantic drift with spoken chapter transitions through cross-modal validation.
- We construct a three-level memory hierarchy (raw video, clip-level memory CLM, event-level memory ELM) atop the detected events, storing content at progressively higher semantic abstraction.
- We design a multi-agent collaborative retrieval framework that queries the hierarchy through query decomposition, tiered search, evidence validation, and answer generation, and show on three benchmarks that the combined system outperforms the strongest baseline by +4.3 / +3.2 / +2.8 points while also reducing retrieval rounds.

2 Related Work

2.1 Long-Context Video Understanding

Applying MLLMs to video typically involves sampling or compressing frames into a single forward pass (Wang et al., 2024; Bai et al., 2025; Li et al., 2024a; Chen et al., 2024). While effective for short clips, this degrades as duration grows because the finite context budget reduces per-frame density. Benchmarks such as Video-MME (Fu et al., 2025) and LVBench (Wang et al., 2025a) quantify this degradation, and EgoLife (Yang et al., 2025) extends it to week-long streams. Two strategies attempt to spend the budget more wisely. Visual token compression (He et al., 2024; Li et al., 2024b; Song et al., 2024) reduces per-frame cost uniformly, sacrificing detail where it matters most. Keyframe selection (Wang et al., 2025d,b) discards low-relevance frames entirely, but a pre-computed relevance score cannot surface brief events in hour-long videos. Both trade fidelity or recall for throughput. Crucially, these methods treat video as a flat frame sequence, ignoring its inherent event structure. Event segmentation has been studied for offline summarization and action recognition, yet rarely as a substrate for interactive memory systems.

2.2 RAG-based Video Understanding

A natural response to the fixed-window limitation is to offload storage to an external memory that is populated offline and queried at inference time. RAG-based approaches index textual captions of fixed-length clips and retrieve the top-k at query time; graph-structured variants like LightRAG (Guo et al., 2025b) encode cross-segment relations to support more structured lookups. Because the memory is built before the question arrives, these systems can scale to arbitrarily long videos without increasing per-query latency.

The benefit comes with a representational cost: visual content is converted to text before storage, so retrieval quality is bounded by what the captioner chose to verbalize. Rare or visually subtle events—the very cases that motivate memory-based designs—are the most likely to be omitted. EgoRAG (Yang et al., 2025) organizes egocentric streams into hourly and daily summaries, but still relies on predetermined temporal scales. SceneRAG (Zeng et al., 2026) builds a scene–entity graph over narrative-consistent scenes, but derives boundaries primarily from ASR transcripts and temporal heuristics, which can be unreliable when speech is sparse or noisy. For RAG systems that retain predetermined temporal units, the atomic memory unit remains a fixed-length segment whose boundaries are chosen before examining content, leaving the system vulnerable to the same boundary effects that plague flat context windows.

2.3 Memory-based Video Understanding

More expressive architectures address representational limits through structured memory. M3-Agent (Long et al., 2025) maintains an entity-centric multimodal graph for multi-turn reasoning. WorldMM (Yeo et al., 2025) combines episodic, semantic, and visual memories with adaptive multi-scale retrieval. VideoLucy (Zuo et al., 2026) constructs a three-level caption hierarchy with agent-driven backtracking. Ego-R1 (Tian et al., 2025) trains an RL agent for dynamic tool calling over week-long recordings. HEM-LLM (Cheng et al., 2025) partitions videos at minima of adjacent-

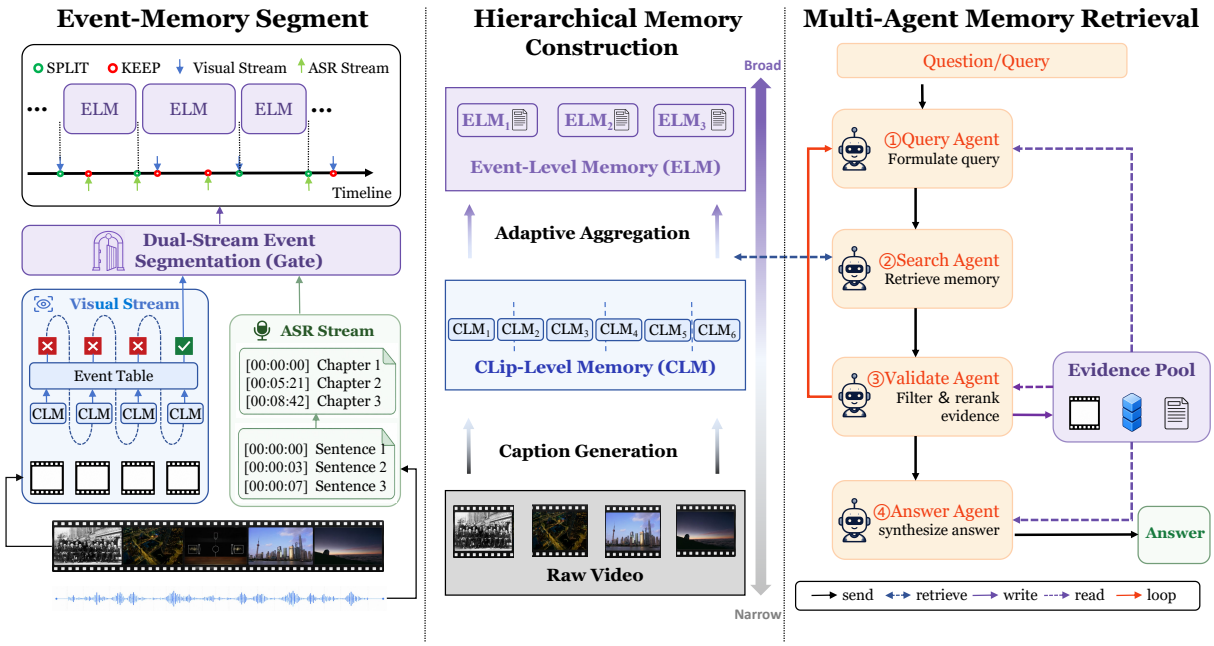


Figure 1: EMMA Framework overview. **Left:** dual-stream event segmentation (visual + ASR) with cross-modal boundary confirmation. **Middle:** three-level memory hierarchy with detected boundaries aggregating CLMs into ELMs. **Right:** four-agent broad-to-narrow retrieval loop.

frame visual similarity and compresses events into local and global visual memories. Video-EM (Wang et al., 2025c) groups query-relevant keyframes into episodic events using visual temporal discontinuities and temporal proximity.

Despite their diversity, memory construction remains a bottleneck. Predefined temporal units can fragment coherent events or dilute brief ones, while visual or visual–text signals can miss speech-led transitions, confuse repeated instances, or omit fleeting cues. Later retrieval must operate over these incomplete units and cannot fully recover the lost event structure.

3 Methodology

3.1 Overview Architecture

We propose EMMA, a multimodal memory framework for long video understanding built around explicit semantic event detection. The system operates in three stages. (1) **Event segmentation:** a dual-stream detector locates semantic event boundaries by fusing visual and linguistic signals through cross-modal validation. (2) **Memory construction:** detected boundaries organize video content into a three-level memory hierarchy of raw video, Clip-level Memory (CLM), and Event-level Memory (ELM). (3) **Memory retrieval:** a multi-agent framework iteratively queries this hierarchy to answer questions. As shown in Figure 1, the pipeline proceeds from left to right: the dual-stream detector locates boundaries; detected boundaries aggregate CLMs into semantically self-contained ELMs; and four specialized agents collaboratively retrieve evidence from coarse events to precise clips. We detail each stage below.

3.2 Dual-Stream Event Segmentation

A core challenge in building semantically coherent memory is deciding where one event ends and another begins. Because a boundary may be marked visually or verbally alone, a single-modality detector is structurally insufficient. We therefore fuse a visual stream tracking semantic drift and an ASR stream proposing chapter-level candidates. Both streams are adjudicated through cross-modal validation guided by an ASR reliability indicator ρ. This design maximizes boundary recall as each stream covers what the other misses, while a one-step look-ahead preserves precision.

Visual Stream. The visual stream maintains a Rolling Event Table (ET), a dynamically updated record that accumulates the semantic narrative of the current event as new clips arrive. Upon receiving a new CLM c_t, an LLM-based updater F_{ET}

refreshes the ET and emits a transition signal:

$$\delta_t = F_{\text{ET}}(\text{ET}_{t-1}, c_t) \in \{\text{CONT, SHIFT}\}, \quad (1)$$

where CONT indicates c_t continues the current event and SHIFT signals a potential boundary. The ET resets upon each confirmed boundary, ensuring it reflects only the ongoing event. A minimum cooldown interval between consecutive splits suppresses spurious visual-only triggers immediately after a confirmed boundary.

ASR Stream. The ASR stream converts the video's audio into structured linguistic signals. ASR transcribes spoken audio into timestamped text transcripts. The resulting transcript $\mathcal{T} = \{(s_i, e_i, w_i)\}_{i=1}^{N}$, where s_i and e_i denote the start and end timestamps and w_i the text content of the i-th utterance, is passed to an LLM producing chapter-level boundary timestamps $\mathcal{B}_{\text{ASR}}$ and an LLM-produced reliability indicator $\rho \in \{\text{HIGH}, \text{LOW}\}$ reflecting whether the ASR transcript contains clear chapter-level structure. The LLM assesses linguistic coherence and the presence of explicit transition cues in the transcript; if the transcript exhibits reliable chapter markers, the LLM sets $\rho = \text{HIGH}$, otherwise LOW. For each CLM c_t spanning $[t_s, t_e)$, the ASR stream emits a boundary indicator:

$$\phi_t = [\exists\, b_k \in \mathcal{B}_{\text{ASR}} \text{ s.t. } t_s \leq b_k < t_e], \quad (2)$$

where $\phi_t = 1$ if a chapter boundary falls within the clip, and 0 otherwise. Both streams propose boundary candidates independently; ρ is forwarded to the validator below as a soft prior on how to weigh their evidence, not as a hard switch.

Cross-Modal Boundary Confirmation. When either stream signals a boundary ($\delta_t = \text{SHIFT}$ or $\phi_t = 1$), the system enters a Pending state: it buffers the current clip c_t and inspects the next clip c_{t+1} before committing. This one-step look-ahead filters transient noise. If c_{t+1} returns to the current event theme, the signal is discarded as noise and the event continues. If c_{t+1} also diverges, a true boundary is likely and the LLM-based validator F_{val} proceeds to adjudication.

The validator jointly analyzes the current event state ET_{t-1}, the two clips c_t and c_{t+1}, and the combined dual-stream evidence from δ_t and ϕ_t to produce a boundary decision v_t and, when applicable, a precise boundary timestamp τ^*:

$$(v_t, \tau^*) = F_{\text{val}}(\text{ET}_{t-1}, c_t, c_{t+1}, \delta_t, \phi_t, \rho). \quad (3)$$

The reliability indicator ρ acts as a soft prior that biases the validator toward the more trustworthy stream: when ρ=HIGH the validator weighs the ASR signal more heavily; when ρ=LOW it relies more on visual evidence. The validator outputs $v_t \in \{\text{KEEP}, \text{SPLIT}\}$. If $v_t = \text{KEEP}$, the signal is rejected as transient noise and the current event continues uninterrupted. If $v_t = \text{SPLIT}$, the validator identifies the utterance in the ASR transcript that marks the new event onset and uses its start timestamp as the precise boundary τ^*, falling back to $t_s(c_t)$ when ASR is unavailable. CLMs preceding τ^* are consolidated into an ELM (§3.4).

3.3 Hierarchical Memory Construction

Based on event boundaries produced by the dual-stream segmentation module, we construct a three-level memory hierarchy comprising raw video, Clip-level Memory (CLM), and Event-level Memory (ELM), each serving a distinct role in balancing detail preservation with semantic abstraction.

Raw Video. Raw video preserves the original uncompressed footage. It is accessed on demand after a relevant time window has been localized through CLM or ELM retrieval, allowing a VLM to recover fine-grained visual details that are lost during caption-based compression.

Clip-level Memory (CLM). CLM is the finest-grained indexed unit. Each CLM corresponds to a τ_{clip}-second video segment with boundaries softly aligned to ASR timestamps. A VLM processes frames at 1 fps and produces:

$$\text{CLM}_t = (S_t, C_t, A_t, [t_s, t_e], \mathbf{e}_t), \quad (4)$$

where S_t is a summary, C_t a visual caption, A_t the ASR transcript, $[t_s, t_e]$ the temporal span, and $\mathbf{e}_t$ the embedding of all fields. The Event Table (§3.2) is fed as context during VLM generation to maintain narrative coherence across clips.

Event-level Memory (ELM). ELM represents a complete semantic event. When a boundary is confirmed, all CLMs since the previous boundary are consolidated into one ELM via LLM summarization:

$$\text{ELM}_k = (S_k, D_k, I_k, [t_s^k, t_e^k]), \quad (5)$$

where S_k is the event summary, D_k a detailed event description, and I_k the reasoning inferred from the event, $[t_s^k, t_e^k]$ the span.

Unlike CLM, which records fragmented observations in fixed windows, ELM distills a more concrete event morphology, enabling the memory retrieval layer to retrieve event-level gist without scanning every CLM.

3.4 Multi-Agent Collaborative Retrieval

A long-video question often requires evidence from multiple temporally distant events, and the right granularity (event gist, clip caption, or raw footage) is unknown in advance. A single retrieval round is insufficient because answers frequently hinge on cross-event reasoning, and a monolithic LLM cannot simultaneously plan, search, and critically validate evidence. We therefore decompose retrieval into an iterative loop shared by four specialized agents.

Query Agent. Plans each round by selecting the target memory level (ELM, CLM, or Raw) and formulating a search query. It follows a broad-to-narrow strategy: early rounds query ELM for coarse event localization; later rounds query CLM or Raw for fine-grained detail.

Search Agent. Executes retrieval against the chosen memory level. For CLM and ELM, it scores each candidate m against query q by combining semantic embedding similarity with keyword-level Jaccard overlap:

$$\text{score}(q, m) = \alpha \cos(\mathbf{e}_q, \mathbf{e}_m) + (1-\alpha)\,\text{Jaccard}(q, m), \tag{6}$$

with weight α. For Raw video, it invokes a VLM to directly observe the localized segment and judge relevance.

Validate Agent. Filters retrieved candidates for relevance, merges them into an evidence pool, and checks whether the accumulated evidence is sufficient to answer the question. If not, it reports missing information to the Query Agent to guide the next round.

Answer Agent. Synthesizes the final answer from the accumulated evidence pool once the validator deems it sufficient, outputting the predicted option with a supporting reasoning trace.

The four agents operate in a shared control loop. The Query Agent plans; the Search Agent executes; the Validate Agent assesses and decides whether to continue; the Answer Agent terminates the loop by producing the final answer. The loop exits when evidence is sufficient or a maximum of rounds is reached.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. We evaluate on three multiple-choice benchmarks that differ in video genre, duration, and question type. Accuracy is reported throughout following official protocols.

- **VideoMME-long** (Fu et al., 2025) is a general long-video understanding benchmark with 300 hour-level videos (avg. 41.4 min) from diverse domains and 900 questions.
- **LVBench** (Wang et al., 2025a) is an adversarial long-content benchmark for structured reasoning, with 103 ultra-long videos (avg. 4,101 s, ∼68 min) and 1,549 questions across six categories: Entity Recognition (ER), Event Understanding (EU), Key Information Retrieval (KIR), Reasoning (Rea), Temporal Grounding (TG), and Summarization (Sum).
- **EgoLifeQA** (Yang et al., 2025) is a first-person persistent-memory benchmark built on week-long egocentric recordings (avg. 44.3 h) with 500 questions on daily activities across five categories: EntityLog (EL), EventRecall (ER), HabitInsight (HI), RelationMap (RM), and TaskMaster (TM).

Baseline Method. To assess EMMA against the full spectrum of existing approaches, we compare four method categories that progress from raw model capacity to structured memory systems. Results for the first three categories are from official leaderboards or original papers; memory-based competitors are re-implemented under our unified backbone (Qwen3-VL-8B + DeepSeek-R1 (Guo et al., 2025a)) to ensure fair comparison.

- **Proprietary models.** We include GPT-4o (Hurst et al., 2024), GPT-5 (Singh et al., 2025), Gemini-1.5-Pro (Team et al., 2024), and Gemini-2.5-Pro (Comanici et al., 2025) as closed-source reference models. Their benchmark results are taken from official leaderboards or prior reports, as summarized in Table 1.
- **Open-source MLLMs.** Qwen2-VL (Wang et al., 2024), LLaVA-OV (Li et al., 2024a),

InternVL2.5 (Chen et al., 2024), and Qwen3-VL-8B (Bai et al., 2025) uniformly sample frames within their context windows. AdaReTaKe (Wang et al., 2025b) reduces visual-token redundancy and VideoChat-Flash (Li et al., 2024b) applies hierarchical visual compression. These methods scale by enlarging context, not by restructuring memory.

- **RAG-based methods.** Video-RAG (Luo et al., 2026) indexes auxiliary text for single-pass retrieval. LightRAG (Guo et al., 2025b) uses a dual-level graph combining entity and summary nodes. EgoRAG (Yang et al., 2025) is the EgoLife baseline. These methods test flat retrieval against our structured multi-round traversal.

- **Memory-based methods.** M3-Agent (Long et al., 2025) (entity graph), VideoLucy (Zuo et al., 2026) (hierarchical caption backtracking), and WorldMM (Yeo et al., 2025) (multi-scale episodic memory). These are our direct competitors that maintain higher-level memory structures without explicit event detection at the base layer.

Our framework EMMA differs from all four categories. Unlike full-context MLLMs, EMMA does not scale by enlarging the context window but by restructuring video into an event-aligned memory hierarchy. Unlike RAG systems, it performs multi-round iterative retrieval across three memory levels rather than single-pass flat retrieval. Unlike existing memory-based agents, it explicitly detects semantic event boundaries at the base layer, making memory units aligned with natural event structure rather than arbitrary temporal windows or entity graphs.

Implementation Details. We use Qwen3-VL-8B (Bai et al., 2025) as the visual encoder for clip captioning and visual observation, sampling video frames at 1 fps; ASR transcripts are extracted from the audio track of each video. DeepSeek-R1 serves as the LLM backbone for all reasoning modules, including Event Table updates, boundary adjudication, ELM summarization, and the four retrieval agents. We set the base clip duration $\tau_{\text{clip}} = 30$ s, the minimum inter-split interval 30 s (i.e., one clip), the hybrid retrieval weight $\alpha = 0.7$, the maximum retrieval rounds $R = 5$, and top-$k = 5$ per round. No task-specific fine-tuning is applied. To verify the contribution of our event-driven memory system, all memory-based competitors (M3-Agent, WorldMM, VideoLucy) are re-implemented under the same Qwen3-VL-8B + DeepSeek-R1 backbone, while their original memory construction and retrieval logic are preserved. Code, prompt templates, and reproduction scripts are available at https://github.com/BAI-LAB/EMMA.

Table 1: Overall accuracy (%) on VideoMME-long, LVBench, and EgoLifeQA. Memory-based systems use a unified Qwen3-VL-8B + DeepSeek-R1 backbone; results for all other methods are taken from official leaderboards or the corresponding papers.

Method	VideoMME-L	LVBench	EgoLifeQA
Proprietary Models			
GPT-4o	65.3	48.9	36.2
GPT-5	74.3	60.4	48.6
Gemini-1.5-Pro	67.4	33.1	36.9
Gemini-2.5-Pro	55.7	57.0	46.4
Open-source MLLMs			
Qwen2-VL-72B	62.2	41.3	–
AdaReTaKe-72B	65.0	53.3	–
LLaVA-OV-72B	60.0	–	30.8
Qwen3-VL-8B	58.4	48.3	38.6
InternVL2.5-78B	62.6	43.6	–
VideoChat-Flash-7B	55.6	–	34.2
RAG-based Systems			
LightRAG	46.6	30.4	48.8
Video-RAG	60.1	33.1	55.4
EgoRAG	41.1	32.2	36.0
Memory-based Systems			
M3-Agent	57.8	51.5	55.6
WorldMM	66.7	54.2	59.6
VideoLucy	63.7	56.3	53.4
EMMA	**71.0**	**59.5**	**62.4**

Table 2: Per-category accuracy (%) on LVBench dataset.

Method	ER	EU	KIR	Rea	TG	Sum	Overall
GPT-4o	48.9	49.5	48.1	50.3	40.9	50.0	48.9
Gemini 1.5 Pro	32.1	30.9	39.3	27.0	31.8	32.8	33.1
Qwen2-VL-72B	38.0	41.1	38.3	46.5	41.4	46.6	41.3
AdaReTaKe-72B	53.0	50.7	62.2	54.7	45.5	37.9	53.3
InternVL2.5-78B	43.8	42.0	42.1	51.0	36.8	37.9	43.6
VideoAgent	28.0	30.3	28.0	28.0	29.3	36.4	29.3
M3-Agent	47.8	52.9	60.2	51.0	47.7	50.9	51.5
WorldMM	48.1	53.5	59.1	50.2	58.2	55.2	54.2
VideoLucy	48.8	59.7	63.9	48.3	50.9	55.2	56.3
EMMA	**56.4**	**60.9**	**64.9**	**54.2**	**60.5**	**56.9**	**59.5**

4.2 Main Results

Table 1 reports overall accuracy on VideoMME-long, LVBench, and EgoLifeQA. For methods whose category-level results are publicly available, Table 2 and Table 3 provide per-category results on LVBench and EgoLifeQA, covering question types from point-level retrieval to event-level reasoning. We make four observations.

Table 3: Per-category accuracy (%) on EgoLifeQA dataset.

Method	EL	ER	HI	RM	TM	Avg.
GPT-4o	34.4	42.1	29.5	30.4	44.4	36.2
Gemini 1.5 Pro	36.0	37.3	45.9	30.4	34.9	36.9
LLaVA-OV-72B	36.8	34.9	31.1	22.4	28.6	30.8
EgoRAG	39.2	36.5	31.1	33.6	39.7	36.0
M3-Agent	46.0	54.8	63.9	59.7	60.3	55.6
VideoLucy	44.4	56.3	59.0	54.8	57.1	53.4
WorldMM	51.6	58.7	**68.9**	58.9	**69.8**	59.6
EMMA	**52.0**	**62.7**	67.2	**66.4**	**69.8**	**62.4**

Table 4: Component ablation on three benchmarks. Acc: accuracy (%); Turn: average retrieval rounds.

Method	VideoMME-L		LVBench		EgoLifeQA	
	Acc	Turn	Acc	Turn	Acc	Turn
Fixed Window	63.6	2.26	57.3	2.95	55.2	3.60
w/o Raw Video	70.0	1.66	59.3	3.39	59.8	3.11
w/o CLM	67.8	1.69	54.4	2.11	54.2	2.94
w/o ELM	67.2	2.03	55.3	2.77	56.8	3.52
w/o Multi-Agent	69.2	–	57.2	–	53.4	–
Full	**71.0**	1.68	**59.5**	2.71	**62.4**	3.16

(1) EMMA consistently outperforms all compared memory-based systems under the unified Qwen3-VL-8B + DeepSeek-R1 setting. Compared with the strongest memory-based baseline on each benchmark, EMMA improves by +4.3, +3.2, and +2.8 points on VideoMME-long, LVBench, and EgoLifeQA, respectively. These gains indicate that event extraction, hierarchical memory, and multi-agent retrieval together yield a stronger design than existing memory organizations. EMMA is also competitive with current proprietary models and achieves the highest unweighted average performance among methods reporting results on all three benchmarks.

(2) Hierarchical memory provides benefits beyond context-window enlargement and caption-based retrieval. Compared with the strongest open-source MLLM on each benchmark, EMMA improves by +6.0, +6.2, and +23.8 points on VideoMME-long, LVBench, and EgoLifeQA, respectively. It also surpasses the strongest RAG-based system by +10.9, +26.4, and +7.0 points. These results suggest that queryable multimodal memory becomes increasingly important as video duration extends, while simply enlarging the context window or retrieving caption-based segments remains insufficient.

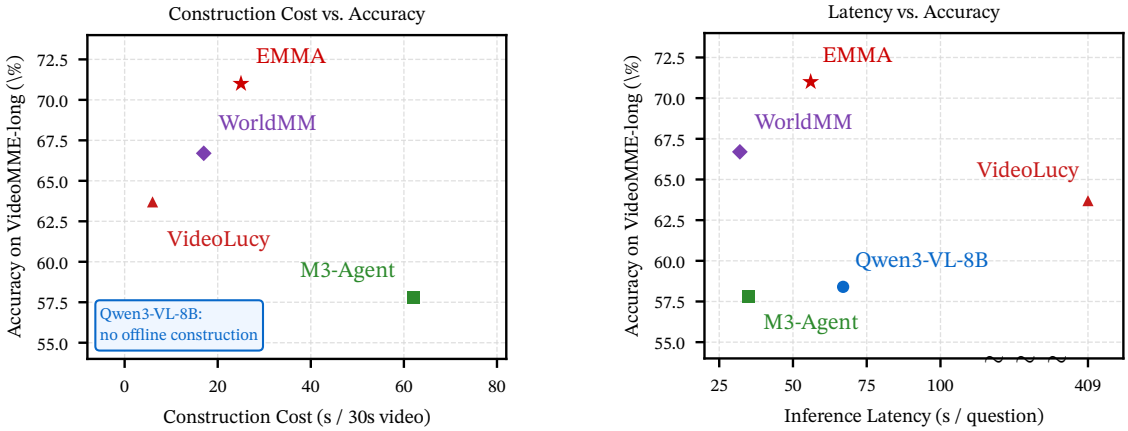
(3) Event-driven memory units strengthen memory organization. Existing memory-based systems maintain higher-level structures, such as entity graphs, hierarchical captions, or multi-scale episodic banks. EMMA additionally constructs its base memory units from detected semantic events. The consistent gains in Table 1, together with the component and segmentation analyses in Section 4.3, suggest that event-aligned cuts are an important factor for memory organization effectiveness when backbone capacity is fixed.

(4) Gains concentrate on event-structured questions. On questions that require reasoning across multiple clips of the same coherent event, as opposed to locating a single moment, EMMA shows consistently positive gains over the strongest memory-based baseline, with the largest reaching +6.7 on EgoLifeQA RM. On question types that only require point-level retrieval (KIR and EL) or pattern aggregation (HI), the effect is mixed. This selectivity is consistent with the improvements arising from the event-alignment mechanism, rather than backbone capacity or incidental data effects.

4.3 Ablation Study

We study EMMA through component ablations and targeted analyses of its event segmentation module.

Component contribution. To verify the contribution of each component in EMMA, we remove one module at a time and measure the resulting drop in accuracy and change in retrieval rounds. From Table 4, we make three observations. (1) Replacing dual-stream event segmentation with fixed-window splitting produces the largest accuracy drop (-7.4, -2.2, and -7.2 points) and requires more retrieval rounds, indicating that event-aligned units provide more retrieval-friendly memory than arbitrary windows. (2) The three memory levels are useful in different ways: removing raw video mostly hurts when direct visual evidence is needed, explaining its limited effect on LVBench; removing CLM loses fine-grained evidence; and removing ELM lowers accuracy while increasing retrieval rounds (2.03 vs. 1.68 on VideoMME-long), showing that ELM provides efficient coarse-level localization. (3) Replacing the multi-agent retrieval loop with a single LLM call decreases performance by -1.8, -2.3, and -9.0 points, respectively, showing the value of iterative planning, search, and validation over a single-call alternative. A complementary single-agent multi-round control is reported in Appendix C.



(a) Accuracy v.s. Memory construction latency.

(b) Accuracy v.s. Inference latency.

Figure 2: Accuracy-efficiency trade-off on VideoMME-long. (a) Offline memory construction latency per 30 s of video. (b) Per-question inference latency.

Event segmentation quality. The Fixed Window ablation above shows the downstream value of event-aligned memory. We further directly evaluate whether EMMA identifies boundaries that agree with external event references. Using creator-provided chapter timestamps from 66 VideoMME-long videos, we report boundary precision, recall, and F1 under 30 s and 60 s tolerances, together with WindowDiff (WD) for overall segmentation consistency. As shown in Table 5, EMMA consistently outperforms fixed-window baselines. Compared with the strongest fixed-window result, EMMA improves F1@30 by 29.2 points and F1@60 by 15.2 points, while achieving the lowest WD. This improvement is also reflected in downstream QA on the same subset, where EMMA reaches 73.23% accuracy compared with 70.71% for Fixed 3min. Detailed evaluation protocols are provided in Appendix C.

Dual-stream segmentation. To isolate the contribution of the two segmentation streams, we evaluate a randomly sampled, fixed one-fifth subset of VideoMME-long (180 questions), while keeping all other components unchanged. Table 6 shows that the dual-stream setting achieves the highest accuracy on this controlled subset. The visual stream captures semantic shifts that may not be verbalized, whereas the ASR stream captures explicit topic transitions. The subset construction and controlled evaluation settings are provided in Appendix C.

4.4 Efficiency Analysis

Although EMMA introduces additional computation through event segmentation, hierarchical memory construction, and multi-round multi-agent retrieval, the cost remains moderate. We evaluate this claim from two aspects. Figure 2a measures offline memory construction cost per 30 s of video: EMMA requires ∼26 s, lower than M3-Agent (∼65 s) and moderately higher than WorldMM (∼15 s). Figure 2b compares per-question inference latency against accuracy: EMMA achieves the highest accuracy (71 %) with a latency of ∼60 s per question, significantly performs Qwen3-VL-8B (∼58 % at similar latency) and substantially faster than VideoLucy, which pays a heavy latency penalty for its iterative backtracking. Memory-based methods such as WorldMM and M3-Agent show lower latency but at the cost of significantly reduced accuracy.

Taken together, EMMA achieves the highest accuracy among the compared memory-based systems while maintaining moderate construction and inference costs. Relative to WorldMM, the fastest memory-based system at inference, EMMA improves accuracy by 4.3 points at the cost of additional latency. This trade-off suggests that the additional computation introduced by hierarchical memory and multi-agent retrieval is worthwhile for long-video understanding.

4.5 Hyperparameter Analysis

The choice of base clip duration τ_{clip} is foundational to the memory hierarchy, as it determines

Table 5: Event segmentation quality on 66 VideoMME-long videos with creator-provided chapter timestamps as external reference boundaries. P/R/F1 are reported under 30 s and 60 s tolerances; lower WD is better.

Method	P@30↑	R@30↑	F1@30↑	P@60↑	R@60↑	F1@60↑	WD↓
Fixed 3min	28.2	32.1	26.8	48.1	57.3	47.4	58.9
Fixed 5min	30.5	19.2	21.1	45.9	30.7	33.5	51.5
Fixed 10min	31.2	9.2	13.0	45.0	14.5	20.3	45.6
EMMA	**57.3**	**67.2**	**56.0**	**64.4**	**74.7**	**62.6**	**38.9**

Table 6: Dual-stream segmentation ablation on a randomly sampled, fixed 180-question subset of VideoMME-long. All non-segmentation components follow the full EMMA configuration.

Segmentation Stream	Accuracy (%)
ASR only	71.11
Visual only	73.33
Visual + ASR	**76.11**

Table 7: Impact of clip duration on LVBench accuracy (%). A 381-question subset covers all six question types.

Question Type	#Q	30s CLM	10s CLM	Δ
Key Info. Retrieval	52	48.1	55.8	+7.7
Summarization	26	50.0	61.5	+11.5
Temporal Grounding	15	60.0	80.0	+20.0
Entity Recognition	123	57.7	58.5	+0.8
Event Understanding	132	61.4	59.8	−1.5
Reasoning	33	57.6	48.5	−9.1
Overall	381	57.2	58.8	+1.6

the granularity of the finest indexed layer. Finer clips provide more detailed captions but increase retrieval complexity; coarser clips reduce the search space but may lose fine-grained evidence. To understand this trade-off, we compare 10 s and 30 s clips on a 381-question LVBench subset covering all six question types.

From Table 7, we can see that retrieval-bound categories (KIR, Sum) benefit from finer clips, improving by +7.7 and +11.5 points respectively, because shorter windows localize evidence more precisely. Conversely, categories requiring event-level integration (EU, Rea) degrade by -1.5 and -9.1 points under the same setting, because overly fine segmentation disrupts the coherence of event narratives. Entity Recognition remains largely unchanged (+0.8). The opposing effects roughly cancel on this subset (+1.6 overall). These results validate a core design choice of our EMMA architecture: no single granularity suits all question types. Our three-level hierarchy resolves this tension naturally. The 30 s default is preferred because holistic-reasoning categories dominate the full LVBench distribution; when fine-grained detail is needed, CLM provides clip-level captions within each event, and Raw video recovers visual specifics on demand. The event-driven ELM layer ensures that finer clips are always retrieved within a coherent semantic unit, avoiding the fragmentation that pure fixed-window splitting would cause.

5 Conclusion

This paper argues that long-video understanding requires memory units aligned with semantic events rather than arbitrary temporal windows. We propose a multi-modal memory system EMMA, a framework that detects event boundaries through cross-modal visual and linguistic streams, organizes memory into a three-level hierarchy of raw video, clips, and events, and retrieves evidence through an iterative multi-agent loop. By grounding memory construction in explicit event segmentation, EMMA ensures that retrieval operates over semantically coherent units rather than fragmented clips. This event-first perspective offers a principled foundation for building richer memory organizations in long-context multimodal systems.

Limitations

EMMA demonstrates that aligning memory units with semantic events through dual-stream segmentation, a three-level hierarchy, and multi-agent retrieval yields substantial accuracy gains over memory-based baselines. Nevertheless, the framework relies on multiple LLM calls during boundary detection and ELM construction, which incurs higher computational overhead than fixed-window approaches. A natural direction is to distill the LLM-based validator into a lightweight trained detector that runs in a single forward pass. Additionally, the base clip granularity and cross-modal boundary confirmation step length are currently

fixed globally; adapting these parameters to local content density would likely close the residual gap further. Although the one-step look-ahead validator suppresses most false boundaries, segmentation errors can still propagate to downstream retrieval, and a formal error-propagation analysis is left to future work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62236003).

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, and 1 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Dingxin Cheng, Mingda Li, Jingyu Liu, Yongxin Guo, Bin Jiang, Qingbin Liu, Xi Chen, and Bo Zhao. 2025. Enhancing long video understanding via hierarchical event-based memory. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

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 *European Conference on Computer Vision*, pages 75–92. Springer.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24108–24118.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025a. Deepseek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2025b. Lightrag: Simple and fast retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13504–13514.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, and 1 others. 2024b. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*.

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

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. *arXiv preprint arXiv:2508.09736*.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and 1 others. 2026. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *Advances in Neural Information Processing Systems*, 38:168008–168033.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, and 1 others. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18221–18232.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

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. *arXiv preprint arXiv:2506.13654*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, and 1 others. 2025a. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967.

Xiao Wang, Qingyi Si, Shiyu Zhu, Jianlong Wu, Li Cao, and Liqiang Nie. 2025b. Adaretake: Adaptive redundancy reduction to perceive longer for video-language understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5417–5432.

Yun Wang, Long Zhang, Jingren Liu, Jiaqi Yan, Zhanjie Zhang, Jiahao Zheng, Ao Ma, Run Ling, Xun Yang, Dapeng Wu, and 1 others. 2025c. Video-em: Event-centric episodic memory for long-form video understanding. *arXiv preprint arXiv:2508.09486*.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, and 1 others. 2025. Egolife: Towards egocentric life assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28885–28900.

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

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.

Nianbo Zeng, Haowen Hou, F Richard Yu, Si Shi, and Ying Tiffany He. 2026. Scenerag: Scene-level retrieval-augmented generation for video understanding. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12122–12126. IEEE.

Jialong Zuo, Yongtai Deng, Lingdong Kong, Jingkang Yang, Rui Jin, Yiwei Zhang, Nong Sang, Liang Pan, Ziwei Liu, and Changxin Gao. 2026. Videolucy: Deep memory backtracking for long video understanding. *Advances in Neural Information Processing Systems*, 38:25660–25691.

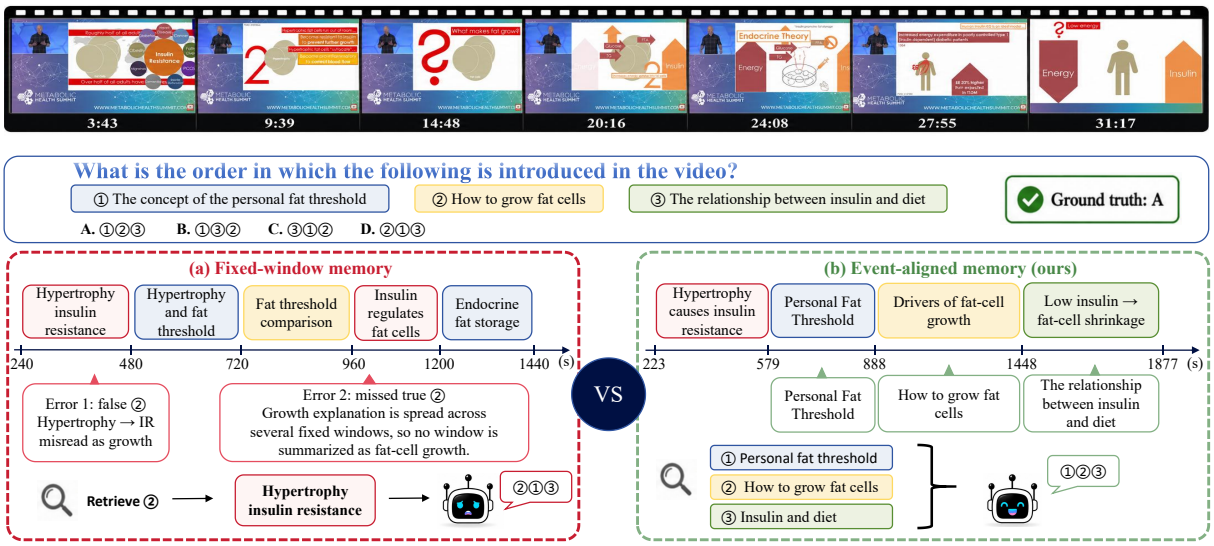


Figure 3: Lecture-video case requiring temporal ordering of three topics. **Left:** fixed-window memory mis-attributes topic ①'s hypertrophy mention to topic ② and fragments topic ②'s true content across three windows, yielding the wrong order. **Right:** event-aligned memory keeps each topic as one self-contained unit and recovers the correct order.

A Case Study

Figure 3 grounds the fragmentation failure described in Section 1 on a lecture-video question that asks for the order of three topics. Two compounding errors trip up the fixed-window system. First, a discussion of hypertrophy that belongs to topic ① ("what we eat → fat") gets misread as evidence for topic ② ("how to grow fat cells"), because hypertrophy is superficially associated with cell growth. Second, the actual content of topic ② is split across three subsequent fixed-windows, so its true onset is hard to localize and the misleading hypertrophy mention dominates retrieval. Event-aligned memory keeps topic ① and topic ② as separate self-contained units, so hypertrophy stays attached to its correct event and topic ②'s onset is recovered in a single retrieval, yielding the correct order.

B Re-implementation of Memory-Based Baselines

To isolate the effect of memory organization, all three memory-based competitors are re-evaluated under the same Qwen3-VL-8B + DeepSeek-R1 backbone used by our framework. We keep each method's memory construction and retrieval logic intact and adapt only the model interfaces.

M3-Agent and VideoLucy. Both systems are originally written against generic VLM/LLM interfaces. We replace their visual encoders with Qwen3-VL-8B for clip captioning and visual observation, and route all reasoning calls (memory updates, retrieval planning, and answer generation) through DeepSeek-R1. No other modifications are made to their public implementations.

WorldMM. WorldMM additionally assumes a QA model that natively ingests retrieved frames alongside text, which DeepSeek-R1 does not support. We therefore adopt the same text-mediated visual interface used by VideoLucy and by our own retrieval loop: when visual input is required, the retrieved visual segment is first passed through Qwen3-VL-8B together with the current question to produce a query-specific description, which is then appended to DeepSeek-R1's context. This adaptation preserves WorldMM's adaptive multi-scale memory selection logic and changes only the visual-to-text bridging layer.

C Evaluation Protocols and Additional Controls

C.1 Event Segmentation Evaluation Protocol

We evaluate event segmentation on the 66 VideoMME-long videos that provide creator-authored chapter timestamps. We treat these timestamps as external reference boundaries; they are used only for evaluation and are not available during memory construction or question answering.

Fixed-window baselines partition each full video at uniform 3, 5, or 10 minute intervals.

We report boundary precision, recall, and F1 under tolerance windows of 30 s and 60 s. These metrics quantify the agreement between predicted and reference boundaries under the corresponding temporal tolerance. We additionally report WindowDiff (WD), which measures disagreement in whether temporally separated points belong to the same segment; lower values indicate better agreement with the reference segmentation. The detailed results are reported in Table 5. For the downstream QA comparison, we restrict evaluation to questions associated with the same 66 videos.

C.2 Controlled Retrieval Studies

For both the dual-stream ablation in Table 6 and the following multi-agent control, we use a randomly sampled and fixed one-fifth subset of VideoMME-long containing 180 questions. All variants use identical questions and keep the remaining memory construction and retrieval settings fixed. The selected question IDs and sampling script are included in the released code repository.

To separate the benefit of role specialization from that of iterative retrieval, we replace EMMA's four specialized agents with a single agent controller while retaining multi-round retrieval. As shown in Table 8, the specialized multi-agent design improves accuracy by 7.22 points while requiring fewer retrieval rounds.

Table 8: Single-agent multi-round control on the fixed 180-question VideoMME-long subset.

Method	Accuracy (%)	Avg. Rounds
Single-agent multi-round	68.89	3.03
EMMA (multi-agent multi-round)	**76.11**	**1.69**