From Verbatim to Gist: Distilling Pyramidal Multimodal Memory via Semantic Information Bottleneck for Long-Horizon Video Agents

Niu Lian[1,2*] **Yuting Wang**[*] **Hanshu Yao**[2] **Jinpeng Wang**[2†]
Bin Chen[2] **Yaowei Wang**[2,3] **Min Zhang**[2] **Shu-Tao Xia**[1]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Harbin Institute of Technology, Shenzhen
[3]Peng Cheng Laboratory
220110904@stu.hit.edu.cn (Niu Lian), wangjp26@gmail.com (Jinpeng Wang)

Abstract

While multimodal large language models have demonstrated impressive short-term reasoning, they struggle with long-horizon video understanding due to limited context windows and static memory mechanisms that fail to mirror human cognitive efficiency. Existing paradigms typically fall into two extremes: vision-centric methods that incur high latency and redundancy through dense visual accumulation, or text-centric approaches that suffer from detail loss and hallucination via aggressive captioning. To bridge this gap, we propose MM-MEM, a pyramidal multimodal memory architecture grounded in *Fuzzy-Trace Theory*. MM-MEM structures memory hierarchically into a *Sensory Buffer*, *Episodic Stream*, and *Symbolic Schema*, enabling the progressive distillation of fine-grained perceptual traces (*verbatim*) into high-level semantic schemas (*gist*). Furthermore, to govern the dynamic construction of memory, we derive a Semantic Information Bottleneck objective and introduce SIB-GRPO to optimize the trade-off between memory compression and task-relevant information retention. In inference, we design an entropy-driven top-down memory retrieval strategy. Extensive experiments across 4 benchmarks confirm that MM-MEM achieves state-of-the-art performance on both offline and streaming tasks, demonstrating robust generalization and validating the effectiveness of cognition-inspired memory organization. Code and associated configurations are publicly available at https://github.com/EliSpectre/MM-Mem.

1 Introduction

To transition from passive observers to truly autonomous agents, intelligent systems require a critical cognitive shift toward long-term and persistent memory (Dong et al., 2025), enabling them to move beyond "here-and-now" perception and interpret continuous, unbounded streams of multimodal information (Qian et al., 2024). While recent Multimodal Large Language Models (MLLMs) (Liu et al., 2023; Li et al., 2022, 2023) have enhanced short-term perceptual modeling (Xie et al., 2024), they lack the efficient memory mechanisms found in human cognition. Specifically, **Fuzzy-Trace Theory** (**FTT**) (Reyna and Brainerd, 1995), a well-established cognitive model, hypothesizes that human memory is not a singular recording but consists of two parallel traces: a *gist trace* that captures abstract semantic meaning and a *verbatim*

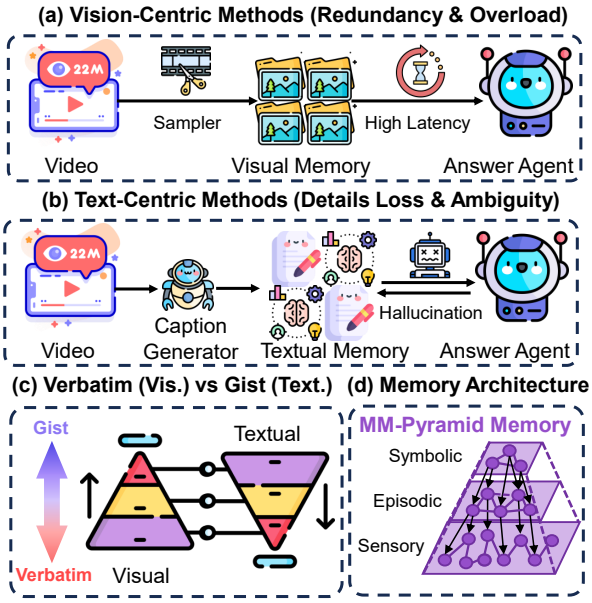


Figure 1: Existing memory paradigms (a-b), inspiration (c), and our insight (d). (a) Vision-centric methods incur redundancy and high latency due to dense visual memories. (b) Text-centric methods suffer from information loss during captioning, leading to hallucination and ambiguity. (c) The natural complementarity between vision and text neatly aligns with the distinction between verbatim and gist traces in *Fuzzy-Trace Theory*. (d) Our MM-MEM is a bottom-up multimodal memory pyramid from sensory buffer to symbolic schema.

[*]These authors contributed equally to this work.
[†]Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11601–11617
July 2-7, 2026

trace that preserves fine-grained perceptual details, allowing to retain specific visual evidence when necessary while efficiently managing long-term semantic context without cognitive overload.

However, existing MLLM-based agents typically fail to strike this biological balance, generally falling into one of two extremes. **Vision-centric paradigms**, such as LongVA (Zhang et al., 2024b) or VideoRAG (Luo et al., 2024), attempt to continuously accumulate visual memories (Fig. 1(a)). While aiming for fidelity, these designs often introduce substantial redundancy due to dense frame sampling. Furthermore, traditional MLLMs (Bai et al., 2025) utilized in these systems often emphasize low-level visual details while overlooking high-level semantic attributes (Long et al., 2025), making it difficult to preserve long-term temporal dependencies. Conversely, **text-centric paradigms** (Shen et al., 2025) convert raw videos into structured textual memories (e.g., knowledge graphs) for efficiency (Fig. 1(b)). Yet, this conversion acts as a lossy compression that discards critical visual cues, leading to ambiguity and hallucinations. Moreover, unlike human memory which is highly dynamic, most existing systems remain static. While dynamic memory management is studied in Large Language Models (LLMs) (Zhong et al., 2024), it is under-explored in multimodal settings. Even recent attempts like A-Mem (Xu et al., 2025) remain text-centric, lacking multimodal grounding for long-horizon reasoning.

To bridge this gap, we propose **MM-MEM**, a novel hierarchical pyramidal multimodal memory architecture inspired directly by the principles of FTT. As illustrated in Fig. 1(c) and (d), MM-MEM is inspired by the complementarity between visual and textual modalities and the distinction between verbatim and gist traces in FTT. Importantly, this connection is realized through cross-modal fusion rather than a rigid one-to-one layer mapping: visual representations predominantly preserve verbatim perceptual evidence, while textual representations mainly encode gist-level semantics. Built upon this principle, MM-MEM spans from perception to cognition across three layers: a *Sensory Buffer* for fine-grained visual evidence, an *Episodic Stream* for event-level summaries, and a *Symbolic Schema* for high-level semantic abstraction. This bottom-up construction progressively transforms perceptual signals into cognitive knowledge.

Crucially, to manage the transition of these layers, we establish a dual mechanism for both adaptive construction and efficient retrieval. For memory construction, we draw motivation from the Information Bottleneck theory (Tishby et al., 2000) and derive a principled formulation to retain maximal semantic content under a constrained budget. Based on this, we introduce **SIB-GRPO** (Semantic-Information Bottleneck GRPO) to optimize the trade-off between semantic preservation and redundancy removal. Complementing this bottom-up construction, we propose an *entropy-driven* top-down memory retrieval strategy. In detail, the agent initiates reasoning at the abstract Symbolic Schema (gist). Only when decision uncertainty is high does the system progressively "drill down" to the Episodic Stream and Sensory Buffer, retrieving fine-grained perceptual traces (verbatim).

To assess the efficacy of this paradigm, we conduct extensive evaluations on 4 challenging benchmarks, covering both offline long-video understanding and online streaming settings. Empirical results demonstrate that MM-MEM not only achieves a new state-of-the-art among open-source MLLMs and agentic systems by notable margins, but also exhibits competitive reasoning capabilities against proprietary models. Qualitative analyses and memory topology visualizations further reveal that MM-MEM can successfully decouple semantic "gist" from visual "verbatim" details, allowing the agent to perform precise detail verification without succumbing to the cognitive overload typical of vision-centric methods. These findings may inspire the development of robust and generalizable cognitive infrastructure for long-horizon autonomous agents.

Our contributions are summarized as follows:

- We propose MM-MEM, a pyramidal multimodal memory architecture grounded in Fuzzy-Trace Theory that bridges the gap between fine-grained perception and high-level cognition.
- We introduce SIB-GRPO, grounded by the Information Bottleneck principle, to optimize the bottom-up memory construction and distill essential knowledge from redundancy.
- We design an entropy-driven top-down memory retrieval strategy that adaptively "drills down" from schemas to details under high uncertainty, ensuring efficient and precise verification.
- Extensive experiments on four benchmarks demonstrate that MM-MEM achieves state-of-the-art performance and robust generalization across both offline and streaming scenarios.

2 Related Work

2.1 Long Video Understanding

While MLLMs have substantially extended vision-language capabilities from images to videos (Liu et al., 2023; Li et al., 2022; Bai et al., 2025; Zhang et al., 2024c), long-video understanding remains constrained by limited context windows. Existing solutions generally fall into two paradigms. *Vision-centric* methods enhance context via dense sampling or token compression (Zhang et al., 2024b; Li et al., 2024; He et al., 2024), with some incorporating auxiliary textual evidence (Luo et al., 2024). Despite improved visual coverage, these approaches often suffer from high computational redundancy and inefficiency. Conversely, *Text-centric* approaches convert videos into captions or structured textual memories for efficiency (Wang et al., 2023, 2025b; Cao et al., 2025; Shen et al., 2025; Wang et al., 2024b). However, such conversion inevitably discards fine-grained visual cues and weakens perceptual grounding, thereby degrading complex reasoning over subtle details.

To bridge this gap, we propose a hierarchical pyramidal multimodal memory that unifies high-level textual memory for coarse localization and low-level visual memory for fine-grained retrieval, achieving a better balance between efficiency and visual fidelity (Gao et al., 2024a).

2.2 Memory for Agents

Memory mechanisms have been widely studied in agents built upon large language models (LLMs). Existing approaches span cache-like hierarchical designs (Packer et al., 2023), forgetting-curve-inspired memory management (Zhong et al., 2024), associative memory linking (Xu et al., 2025), and reinforcement learning-based memory control (Yan et al., 2025). However, these systems, including LicoMemory (Huang et al., 2025), remain largely *text-centric*, limiting their ability to align information across modalities and to preserve the rich structural properties of real-world experiences.

In contrast, memory systems for *multimodal agents* remain relatively underexplored. Prior work such as M3-Agent (Long et al., 2025) typically relies on predefined memory structures and fixed operational workflows, which may constrain generalization in open-ended, long-horizon environments. These limitations highlight the need for a flexible and generalizable memory framework that can support long-term multimodal interactions.

3 Method

We propose MM-MEM, a multimodal memory architecture that helps agents perceive and understand the external world. MM-MEM uses a bottom-up, offline-built hierarchical memory pyramid, spanning from a perception-level *Sensory Buffer* that retains fine-grained visual evidence to a semantics-level *Symbolic Schema* that stores high-level textual abstractions. For long-horizon interaction, we introduce SIB-GRPO for dynamic memory management, which removes redundant memories while preserving task-relevant semantics. We further design a top-down hierarchical retrieval mechanism guided by predictive entropy, which adaptively selects the retrieval depth to balance evidence coverage and resource constraints.

3.1 Multimodal Pyramid Memory Structure

Rather than physically decoupling modalities into isolated tracks, our three-layer hierarchy maintains integrated multimodal representations across all levels. While we conceptually map visual data to verbatim traces and textual data to gist traces inspired by FTT, both modalities coexist throughout the pyramid. The memory construction evolves through a representational shift: it begins as a vision-dominant multimodal memory at the bottom for fine-grained perception, and progressively distills into a text-dominant representation at the top for high-level cognitive abstraction.

Sensory Buffer. Given a long video stream $\mathcal{V}$, we apply content-adaptive temporal segmentation to obtain clips $\mathcal{C} = \{c_t\}$ (e.g., PySceneDetect). For each clip c_t, we identify salient temporal indices $\mathcal{S}_t$ based on inter-frame variation and construct short key sub-clips centered at these indices (details in Appendix C). Sensory memory is instantiated as

$$\mathcal{M}_{\text{sens}} = \{(\mathbf{v}_{t,i}, \mathbf{l}_{t,i}, \tau_{t,i}) \mid i \in \mathcal{S}_t,\ c_t \in \mathcal{C}\}, \quad (1)$$

where $\mathbf{v}_{t,i}$ is the visual representation, $\mathbf{l}_{t,i}$ is the associated text trace (e.g., subtitles or clip captioning), and $\tau_{t,i}$ is the temporal location (e.g., center-frame timestamp). Crucially, at this vision-dominant foundational layer, the text trace $\mathbf{l}_{t,i}$ is strongly bound to its visual counterpart and functions as an auxiliary component. Rather than acting as an independent or high-level semantic representation, the text serves purely as a descriptive label for visual entities. It provides a semantic anchor to help index and isolate the dense, highly redundant verbatim visual details.

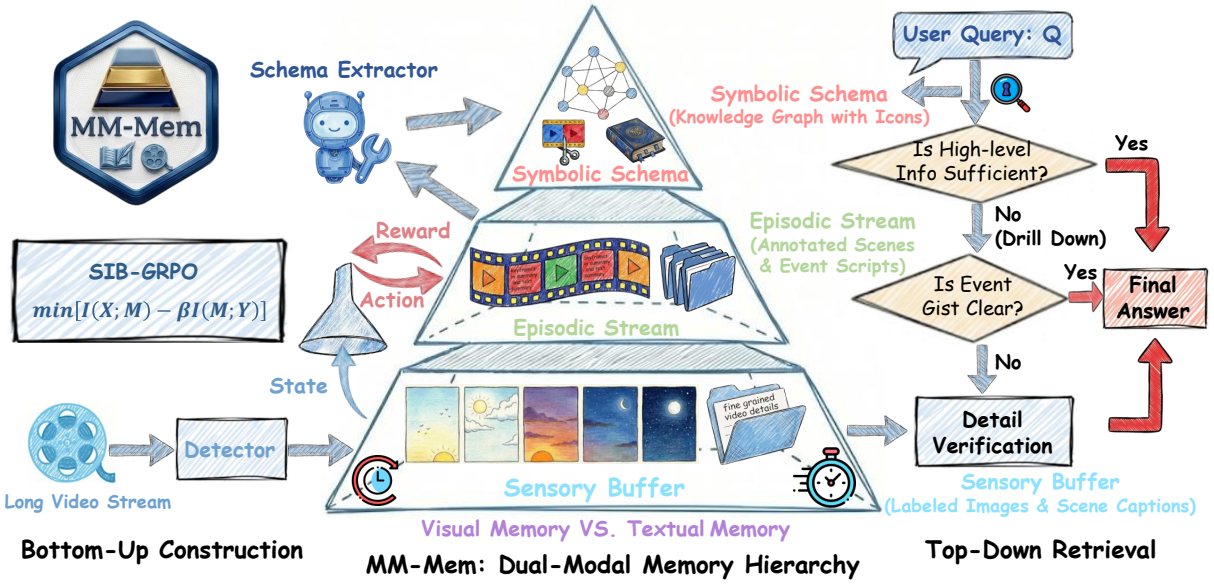


Figure 2: Overview of MM-MEM: MM-MEM unifies visual and textual memory through (left) a bottom-up memory construction process, which transforms raw sensory frames into abstract symbolic schemas, and (right) a top-down retrieval process that supports query-adaptive reasoning.

Episodic Stream. Following selective encoding and temporal contiguity (Dong et al., 2025), we construct an *Episodic Stream* by consolidating sensory entries from $\mathcal{M}_{\text{sens}}$. Each sensory item is $m_{t,i} = (\mathbf{v}_{t,i}, \mathbf{l}_{t,i}, \tau_{t,i}) \in \mathcal{M}_{\text{sens}}$. We maintain an ordered episodic sequence $\mathcal{M}_{\text{epi}}$, with the latest retained node $e^\star = (\mathbf{e}^\star, \mathbf{l}^\star, \tau^\star)$.

For each $m_{t,i}$, a decision operator $\psi(\cdot)$ updates the stream:

$$a_{t,i} = \psi(m_{t,i}, e^\star), \qquad a_{t,i} \in \mathcal{O}, \tag{2}$$

where $\mathcal{O} = \{\text{ADD\_NEW}, \text{MERGE}, \text{DISCARD}\}$. ADD_NEW appends a node initialized from $m_{t,i}$; MERGE integrates $m_{t,i}$ into $e^\star$; and DISCARD removes redundant or low-novelty items. A single chronological pass over $\mathcal{M}_{\text{sens}}$ yields a compact episodic stream.

To obtain event-level abstractions, we cluster retained visual representations (e.g., via K-means) and select representative prototypes as summaries. The resulting episodic memory is

$$\mathcal{M}_{\text{epi}} = \{(\mathbf{e}_k, \mathbf{l}_k, \tau_k)\}_{k=1}^{|\mathcal{M}_{\text{epi}}|}, \tag{3}$$

where $\mathbf{e}_k$ is the representation of the k-th episodic unit, $\mathbf{l}_k$ aggregates its associated textual traces, and τ_k records its temporal span.

Symbolic Schema. To support cross-episode reasoning (Dong et al., 2025), we build a *Symbolic Schema* as a knowledge graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ over episodic memory $\mathcal{M}_{\text{epi}}$. An MLLM extracts entities and glosses from each episodic unit $e_k \in \mathcal{M}_{\text{epi}}$, unifying them into a global prototype set $\mathcal{U}$ with aggregated glosses t_u.

The graph nodes $\mathcal{N} = \{v_k\} \cup \mathcal{U}$ comprise episodic units and prototypes. The edges $\mathcal{E}$ contain optional semantic relations $u_p \xrightarrow{r} u_q$ and, crucially, grounding edges (v_k, u). These grounding edges serve as explicit multimodal pointers: rather than collapsing into a unimodal textual graph, they tightly anchor text-dominant concepts (semantic gist) back to specific episodic units (retaining verbatim visual evidence).

Thus, symbolic memory is instantiated as a text-driven multimodal index:

$$\mathcal{M}_{\text{sym}} = \{(u, t_u, \mathcal{P}_u)\}_{u \in \mathcal{U}}, \tag{4}$$

where $\mathcal{P}_u = \{v_k \mid (v_k, u) \in \mathcal{E}\}$ denotes the visual pointers for concept u. This design enables efficient high-level cognitive reasoning while preserving dynamic drill-down to concrete visual details.

3.2 Bottom-Up Memory Construction

A bottom-up pipeline transforms raw videos into a three-level memory hierarchy: *Sensory Buffer*, *Episodic Stream*, and *Symbolic Schema*. Fine-grained perceptual signals are retained in the *Sensory Buffer*; segments are organized into compact event traces in the *Episodic Stream*; and structured knowledge is consolidated in the *Symbolic Schema* over longer time scales.

During Sensory-to-Episodic construction, redundancy must be compressed while preserving task-relevant semantics and controlling memory growth. We introduce SIB-GRPO to fine-tune the memory manager with reinforcement learning, enabling adaptive generation of information-dense episodic traces. Crucially, driven by SIB-GRPO, this bottom-up construction is not merely a process of data reduction; it orchestrates a smooth transition in modality dominance. The system progressively distills high-information-density textual gist from the heavily redundant visual verbatim traces, naturally aligning with the principles of Fuzzy-Trace Theory. We next describe the Sensory-to-Episodic pipeline and optimization.

3.2.1 Sensory-to-Episodic Memory

Given a sensory buffer $\mathcal{M}_{\text{sens}} = \{m_{t,i} = (\mathbf{v}_{t,i}, \mathbf{l}_{t,i}, \tau_{t,i})\}$, we construct a compact episodic stream $\mathcal{M}_{\text{epi}}$ that retains task-relevant semantics for downstream reasoning while discarding redundant, low-novelty details. Let X denote the sensory memory content (a temporally local window from $\mathcal{M}_{\text{sens}}$ together with the latest episodic node $e^\star$), and let M denote the episodic representation produced by a stochastic encoder (memory manager) $p_\theta(m \mid x)$. We cast this *Sensory-to-Episodic* conversion as stochastic compression, where a memory manager serves as an encoder mapping sensory X to episodic memory M.

Remark: Action-output correspondence. Under fixed update rules, $\mathcal{O}$ uniquely determines M (see Appendix A.3).

Semantic IB Formulation. Let X denote the sensory memory content (a temporally local window from $\mathcal{M}_{\text{sens}}$ together with the latest episodic node $e^\star$), and let M denote the episodic representation produced by a stochastic encoder (memory manager) $p_\theta(m \mid x)$. We adopt an Information Bottleneck (IB) objective (Tishby et al., 2000):

$$\min_{p_\theta(m|x)} \big[I(X;M) - \beta\, I(M;Y)\big], \tag{5}$$

where Y is the supervision label (ground-truth VQA answer) and β controls the compression–prediction trade-off. Here $I(\cdot;\cdot)$ denotes the mutual information between two random variables.

To obtain a tractable training objective, we introduce a variational decoder $q_\phi(y \mid m)$ to approximate $p_\theta(y \mid m)$ and a variational prior $r(m)$ to approximate the marginal $p_\theta(m)$. We define

$$\mathcal{L}_{\text{p}}(\theta,\phi) \triangleq \mathbb{E}_{p(x,y)\,p_\theta(m|x)}\big[\log q_\phi(y \mid m)\big], \tag{6}$$

$$\mathcal{L}_{\text{c}}(\theta) \triangleq \mathbb{E}_{p(x)}\Big[D_{\text{KL}}\big(p_\theta(m \mid x) \,\|\, r(m)\big)\Big]. \tag{7}$$

As shown in Appendix A, $\mathcal{L}_{\text{p}}$ lower-bounds $I(M;Y)$ up to an additive constant $H(Y)$, and $\mathcal{L}_{\text{c}}$ upper-bounds $I(X;M)$. Therefore, dropping the constant $H(Y)$, we optimize the following variational IB objective:

$$\max_{\theta,\phi}\ \beta\,\mathcal{L}_{\text{p}}(\theta,\phi)\ -\ \mathcal{L}_{\text{c}}(\theta). \tag{8}$$

A Quality–Quantity Prior. To encode an explicit *quality–quantity* trade-off in episodic memory, we adopt the prior

$$r(m)\ \propto\ \underbrace{p_{\text{ref}}(m)}_{\text{Quality}}\cdot\underbrace{e^{-\lambda|m|}}_{\text{Quantity}}, \tag{9}$$

where λ is a hyperparameter, $|m|$ is the token length of the textual trace stored at an *Episodic Stream* node, and p_{ref} is a *teacher* distribution promoting fluent, general-purpose memory expressions while anchoring the policy to a trusted prior. This factorization yields an IB regularizer that decomposes into a length penalty and a KL term to the teacher, resembling RLHF-style trust-region objectives.

SIB-GRPO: Dynamic Management. While the IB objective provides a principled semantic compression criterion, the episodic trace is generated discretely (LLM-style) and must be optimized with sequence-level feedback. We therefore train the Memory Manager as a policy $\pi_\theta(m \mid s)$ using reinforcement learning, where $s = (x, M_{\text{old}})$ and the action is a textual trace m appended to $\mathcal{M}_{\text{epi}}$.

For each sampled trace m, we compute the designed scalar reward

$$\begin{aligned} r(s,m) \triangleq\ & \underbrace{R_{\text{vqa}}(s,m)}_{\text{Task Reward}}\ -\ \underbrace{\beta_1\cdot\text{Length}(m)}_{\text{from } e^{-\lambda|m|}} \\ & -\ \underbrace{\beta_2\cdot\log\frac{\pi_{\theta_{\text{old}}}(m \mid s)}{\pi_{\text{ref}}(m \mid s)}}_{\text{from } p_{\text{ref}}(m)}. \end{aligned} \tag{10}$$

Here π_{ref} (equivalently p_{ref}) is a fixed reference policy/distribution. In practice, the log-ratio is evaluated on sampled m from the behavior policy $\pi_{\theta_{\text{old}}}$ and serves as a KL-style regularizer that anchors π_θ to π_{ref}.

Given a state $s = (x, M_{\text{old}})$, we sample a group of G candidate episodic traces $\{m_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot \mid s)$, compute their scalar rewards $\{r_i\}_{i=1}^G$, and construct a standardized group-relative advantage A_i (computed by normalizing each r_i within the

group). We then optimize a PPO-style clipped surrogate using the importance ratio $\rho_i(\theta) = \pi_\theta(m_i \mid s)/\pi_{\theta_{\text{old}}}(m_i \mid s)$, where ϵ controls the clipping range. The resulting SIB-GRPO objective is

$$J_{\text{SIB-GRPO}}(\theta) = \mathbb{E}_{s,\{m_i\}\sim\pi_{\theta_{\text{old}}}}\left[\frac{1}{G}\sum_{i=1}^{G}\min\left(\rho_i(\theta)A_i,\ \text{clip}\big(\rho_i(\theta),\,1-\epsilon,\,1+\epsilon\big)\,A_i\right)\right], \tag{11}$$

In practice, we minimize $-J_{\text{SIB-GRPO}}(\theta)$ as the training loss.

3.3 Entropy-Driven Top-Down Retrieval

A top-down, coarse-to-fine retrieval strategy is adopted, querying memory from high-level semantic abstractions to progressively finer perceptual evidence. Retrieval begins at the *Symbolic Schema* level via text to rapidly instantiate the semantic gist. If uncertainty persists, the query descends to lower layers and may terminate at the *Episodic* layer once sufficient evidence is obtained; only when ambiguity remains is it routed to the *Sensory Buffer*, retrieving CLIP-encoded visual keyframes to resolve decisions with local visual details.

This design follows Reverse Hierarchy Theory (Hochstein and Ahissar, 2002), which posits that perception is initiated with high-level *vision at a glance* and refined by low-level *vision with scrutiny* when fine-grained discrimination is required.

Formally, given a question $\mathcal{Q}$ and a candidate answer set $\mathcal{A} = \{a_1, \ldots, a_N\}$, each retrieval step s returns evidence R_s. Given accumulated evidence $R_{\le s}$, a posterior distribution $p(a_i \mid \mathcal{Q}, R_{\le s})$ is maintained over candidates. For notational convenience, let

$$p_i^{(s)} \triangleq p(a_i \mid \mathcal{Q}, R_{\le s}). \tag{12}$$

The entropy of this answer distribution is used as an adaptive stopping criterion:

$$H_s(\mathcal{Q}) = -\sum_{i=1}^{N} p_i^{(s)} \log p_i^{(s)}. \tag{13}$$

Retrieval is terminated when $H_s(\mathcal{Q}) \le \gamma$, or when the entropy reduction $\Delta H_s = H_{s-1}(\mathcal{Q}) - H_s(\mathcal{Q})$ falls below a small ϵ for several consecutive steps, and the most probable answer is returned: $\arg\max_{a_i\in\mathcal{A}} p_i^{(s)}$. Intuitively, rapid semantic narrowing of $\mathcal{A}$ is enabled by high-level text retrieval, whereas low-level keyframe retrieval is invoked only under high entropy, yielding a compute–accuracy trade-off that adapts to question difficulty.

4 Experiment

4.1 Experimental Setup

Benchmarks. Four benchmarks are comprehensively evaluated. **(i) Standard long-video datasets.** VIDEO-MME (Fu et al., 2025a) covers three length regimes, namely short (< 2 min), medium (4–15 min), and long (30–60 min), with 300 videos each (900 total) and 2,700 questions; we report both *with-subtitle* and *without-subtitle* settings. MLVU (Zhou et al., 2025) **dev set** includes nine tasks with videos from 3 min to 2 h (avg. ∼12 min). **(ii) Standard online streaming dataset.** VSTREAM-QA (Zhang et al., 2024a) contains VSTREAM-QA-EGO and VSTREAM-QA-MOVIE for egocentric and third-person narrative understanding. **(iii) A derived egocentric long-video dataset.** We built HD-EPIC++ from HD-EPIC (Perrett et al., 2025) by re-splitting train/test, comprising 156 videos; details see Appendix B.1.

Evaluation protocol. For comparison, we follow baseline settings. For VIDEO-MME, MLVU, and HD-EPIC++, accuracy is used as the evaluation metric. Since VSTREAM-QA consists of open-ended questions, GPT-4O-MINI (Hurst et al., 2024) is leveraged as an automatic judge. We report both accuracy and the averaged score on VSTREAM-QA. Implementation details and evaluation scripts are provided in Appendix D.

Implementation details. Experiments are conducted on NVIDIA A100 80GB GPUs. MM-MEM uses Qwen3-VL-8B (Bai et al., 2025) as the base model. For text retrieval, we use BGE-LARGE-EN-V1.5 (Xiao et al., 2024) and BGE-RERANKER-V2-M3 (Multi-Granularity, 2024). For visual retrieval, we use *clip-level* retrieval by jointly scoring keyframes per clip; CLIP-style embeddings come from the base model's vision encoder. Models are served with vLLM, and fine-tuning is performed under SWIFT with SIB-GRPO. We set $\beta_1 = 0.1$, $\beta_2 = 0.3$, and temperature to 0.0. Hyperparameters are provided in Appendix D.

4.2 Comparison with State-of-the-arts

Long Video Understanding. Baselines follow each benchmark leaderboard and common protocols in prior work, covering (a) proprietary multimodal models, (b) open-source MLLMs, and (c) agent-based systems for long-horizon video understanding. As shown in Table 1, MM-MEM consistently outperforms prior agent systems. Compared with the strongest agent baseline VGENT,

Method	Video-MME								MLVU
	Short		Medium		Long		Overall		M-Avg
	w/o	w/	w/o	w/	w/o	w/	w/o	w/	
Proprietary Models									
Gemini 1.5 Pro (Team et al., 2024)	81.7	84.5	74.3	81.0	67.4	77.4	75.0	81.3	-
GPT-4o (Hurst et al., 2024)	80.0	82.8	70.3	76.6	65.3	72.1	71.9	77.2	64.6
Gemini 1.5 Flash (Team et al., 2024)	78.8	79.8	68.8	74.7	61.1	68.8	70.3	75.0	-
GPT-4V (Achiam et al., 2023)	70.5	73.2	55.8	59.7	53.5	56.9	59.9	63.3	49.2
Open-Sourced MLLMs									
Qwen2-VL-72B (Wang et al., 2024a)	80.1	82.2	71.3	76.8	62.2	74.3	71.2	77.8	-
LLaVA-Video-72B (Zhang et al., 2024c)	81.4	82.8	68.9	75.6	61.5	72.5	70.6	76.9	73.1
Qwen3-VL-8B (Bai et al., 2025)	76.4	79.2	62.3	72.6	55.9	70.8	64.9	74.2	65.9
VideoLLaMA 3-7B (Zhang et al., 2025)	80.1	80.2	63.7	69.6	54.9	61.0	66.2	70.3	73.0
VideoLLaMA 2-72B (Cheng et al., 2024)	69.8	72.0	59.9	63.0	57.6	59.0	62.4	64.7	-
VITA 1.5-7B (Fu et al., 2025b)	67.0	69.9	54.2	55.7	47.1	50.4	56.1	58.7	60.4
Long-LLaVA-7B (Wang et al., 2024c)	61.9	66.2	51.4	54.7	45.4	50.3	52.9	57.1	-
LongVA-7B (Zhang et al., 2024b)	61.1	61.6	50.4	53.6	46.2	47.6	52.6	54.3	-
Video-LLaVA-7B (Lin et al., 2024)	45.3	46.1	38.0	40.7	36.2	38.1	39.9	41.6	-
Agent-based Systems									
Vgent (Shen et al., 2025)	-	-	-	-	-	-	68.9	74.3	72.1
VideoRAG (Luo et al., 2024)	-	67.1	-	60.4	-	60.1	60.5	62.6	72.4
VideoMiner (Cao et al., 2025)	65.6	-	57.5	-	52.2	-	58.4	-	65.1
VideoTree (Wang et al., 2025b)	55.5	-	49.2	-	39.3	-	48.0	-	60.4
MM-MEM (Ours)	**81.5**	**82.8**	**69.6**	**75.8**	**66.1**	**75.7**	**72.4**	**78.1**	**77.2**

Table 1: Comparison on two long-video understanding benchmarks: **VIDEO-MME** and **MLVU**. For VIDEO-MME, we report results under both w/ and w/o subtitle settings (w/ = with subtitles; w/o = without subtitles). For MLVU, we report M-Avg. Most baseline results are taken from the official leaderboards (as of 2026-01-01) or the respective papers. Results not reported in the original sources are marked with "–". For agent-based systems, we report the best-performing configuration with a comparable parameter scale.

Method	VStream-QA-Ego	
	Accuracy	Score
Video-ChatGPT (Maaz et al., 2024)	51.7	3.7
MovieChat (Song et al., 2024)	52.2	3.4
Chat-UniVi (Jin et al., 2024)	50.9	3.8
LLaMA-VID (Li et al., 2024)	54.8	3.9
Flash-VStream (Zhang et al., 2024a)	59.0	3.9
MM-MEM (Ours)	**62.5**	**4.1**

Table 2: Test performance on **VSTREAM-QA-EGO**.

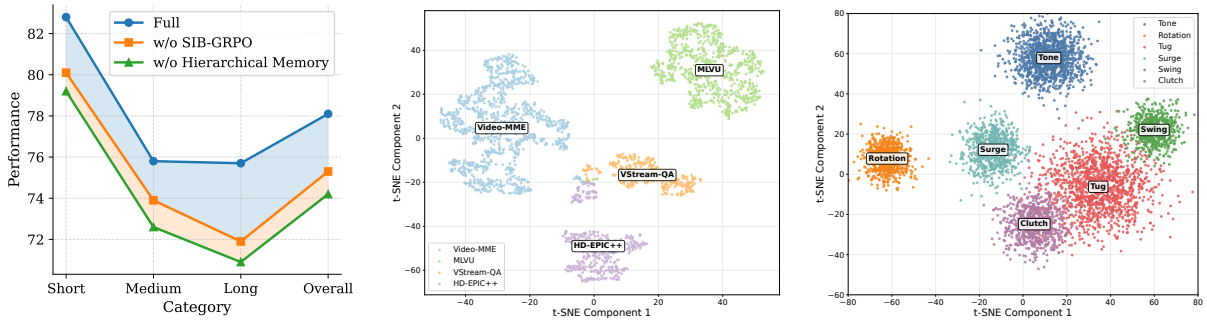
Method	HD-EPIC++
	Accuracy
Qwen3-VL-8B (Bai et al., 2025)	25.88
Qwen2.5-VL-7B (Bai et al., 2025)	24.37
LLaVA-Video-7B (Bai et al., 2025)	25.37
VideoLLaMA 3-7B (Zhang et al., 2025)	20.36
Qwen3-VL-4B (Bai et al., 2025)	24.91
Qwen3-VL-2B (Bai et al., 2025)	22.80
MM-MEM (Ours)	**30.28**

Table 3: Evaluation on the built **HD-EPIC++**.

MM-MEM yields a **5.1**% *relative* gain on VIDEO-MME (both w/o- and w/-subtitle) and a **7.1**% gain on MLVU in **M-Avg**. Despite using QWEN3-VL-8B as the backbone, MM-MEM surpasses all compared open-source MLLMs (e.g., QWEN2-VL-72B) and is competitive with strong proprietary models such as GEMINI 1.5 PRO.

Online Streaming Video Understanding. Unlike most prior work that evaluates only on long-video benchmarks, long-video understanding primarily targets an offline setting, where the model is provided with the user query and a single video clip of finite length at the same time. To better approximate real-world online video-stream scenarios, we additionally evaluate on the streaming benchmark VSTREAM-QA-EGO. As shown in Table 2, MM-MEM remains effective for long-horizon streaming inputs, improving over the previous best method FLASH-VSTREAM by **5.9**% and **5.2**% in terms of **Accuracy** and **Score**, respectively.

Egocentric Long Video Understanding. Table 3 reports accuracy on HD-EPIC++. Our MM-MEM



(a) Ablation results. (b) Sensory Buffer Visual Memory. (c) Episodic Stream Visual Memory

Figure 3: Visualization of ablation results and memory representations.

achieves **30.28%**, outperforming all baselines. It exceeds the strongest competitor (Qwen3-VL-8B) by **+4.40** points (30.28 vs. 25.88), and also surpasses LLaVA-Video-7B and VideoLLaMA 3-7B by **+4.91** and **+9.92** points. This suggests MM-MEM better aggregates fine-grained egocentric cues over long temporal contexts.

4.3 Ablation Studies

Effectiveness of SIB-GRPO and Pyramid Memory. Figure 3a reports an ablation study over *Short*, *Medium*, *Long*, and *Overall* splits. Our full model performs best across all categories, indicating positive contributions from each component. Removing SIB-GRPO consistently degrades performance, with the largest drop on *Long*, suggesting its importance for consolidation under long temporal dependencies. Further removing the Pyramid (hierarchical) memory yields an additional decrease, again most pronounced on *Long* and *Overall*. These results show that pyramid memory complements SIB-GRPO by organizing information at multiple temporal/semantic granularities, improving retention and retrieval for long-horizon reasoning. Additional ablations and hyper-parameter analyses are provided in Appendix E.

Topology of the Cognitive Memory Space. To qualitatively assess the structure of our hierarchical memory, we project memory embeddings to 2D with t-SNE. Figure 3b (Middle) shows Sensory Buffer representations across benchmarks: the clear separation between egocentric (HD-EPIC++) and cinematic (Video-MME) domains indicates that the L1 layer preserves domain-specific visual details without collapse. Figure 3c (Right) visualizes the Episodic Stream, where semantic clusters emerge naturally (e.g., 'Rotation' vs. 'Swing'), suggesting that RL-driven consolidation suppresses noise and promotes abstraction from sensory signals to higher-level reasoning.

Method	VRAM(GB)	Con. Time(s)	Infer. Latency(s)
Proprietary MLLMs			
VideoAgent	N/A	N/A	67.25
Open-Source MLLMs			
Qwen3-VL-8B	22.8	N/A	6.47
Video-RAG	23.0	N/A	25.93
Vgent	18.7	20.18	7.38
MM-Mem (Ours)	**17.8**	**19.54**	**5.35**

Table 4: Efficiency comparison on Video-MME without subtitles. **Con. Time** denotes construction time, and **Infer. Latency** denotes inference latency. All timing metrics are normalized to the time required to process one minute of video.

4.4 Efficiency and Deployment Analysis

We analyze the efficiency of MM-MEM on **Video-MME** without subtitles, covering the Short, Medium, and Long splits. Since runtime is highly correlated with raw video duration, we normalize all timing metrics by the *time required to process one minute of video*, ensuring fair and comparable evaluation. We report **Peak VRAM**, **Memory Construction Time**, and **Inference Latency**. Here, **N/A** denotes results that we are unable to reproduce due to closed-source implementations or insufficient implementation details.

4.5 Qualitative Analysis

As shown in Table 4, MM-MEM achieves a favorable trade-off among construction cost, online latency, and GPU memory usage. Its offline memory construction can be amortized across multiple queries for the same video, which is especially beneficial in Video-MME-style settings. Meanwhile, MM-MEM supports efficient online inference, requiring only **5.35s** per minute of video, with **19.54s** per minute for memory construction. Moreover,

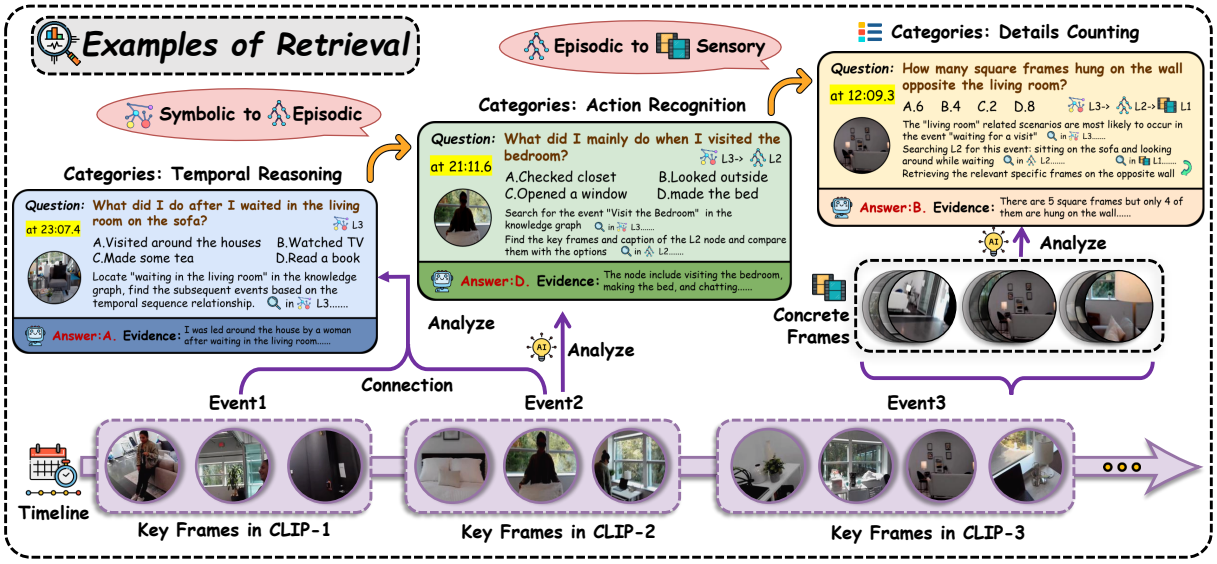


Figure 4: A qualitative example of MM-MEM's coarse-to-fine retrieval across memory layers.

MM-MEM reduces deployment cost by relying on compact high-level textual memory, reaching only **17.8 GB** peak VRAM on an NVIDIA A100, lower than Qwen3-VL-8B and Video-RAG.

To provide more intuitive insights, we present several representative examples in Figure 4. For *Temporal Reasoning*, MM-MEM primarily operates over knowledge graph in Symbolic Schema, where high level semantic abstractions and relational structure allow it to recover the temporal order of actions without directly revisiting low level visual details. For *Action Recognition*, MM-MEM further drills down into Episodic Stream, where temporally localized visual evidence, aligned with event-level summaries, enables the recognition of more fine-grained actions in the current scene. For detail-sensitive tasks such as *Details Counting*, MM-MEM descends all the way to Sensory Buffer, where concrete visual cues at a finer granularity can be retrieved to support precise verification.

Overall, these examples illustrate that the pyramidal multimodal memory of MM-MEM supports a coarse-to-fine retrieval process across memory layers, allowing the model to progressively move from abstract semantic reasoning to detailed perceptual verification as task demands increase. This design not only improves prediction accuracy, but also effectively exploits the complementary strengths of textual and visual representations.

5 Conclusion

In this work, we present MM-MEM, a pyramidal multimodal memory framework grounded in Fuzzy-Trace Theory. By structurally decoupling verbatim visual details from gist semantic schemas, MM-MEM effectively bridges the gap between fine-grained perception and high-level cognition. To govern memory construction, we propose SIB-GRPO, an information-theoretic approach for dynamic, redundancy-aware compression. Complementing this, we introduce an entropy-driven top-down retrieval strategy that adaptively drills down from abstract symbolic schemas to fine-grained sensory details under high uncertainty, ensuring efficient and precise information access. Extensive experiments demonstrate that MM-MEM achieves state-of-the-art performance and robust generalization, inspiring foundational cognitive infrastructure for long-horizon autonomous agents.

Limitations

While MM-MEM demonstrates robust performance in long-horizon video understanding, we identify several avenues for future optimization and research. (i) *Computational Overhead vs. Reasoning Depth:* Our hierarchical architecture prioritizes precise, multi-granularity reasoning, which naturally incurs a higher computational cost during the construction phase compared to flat, compression-heavy models. However, the modular nature of the Sensory, Episodic, and Symbolic layers allows for asynchronous processing and parallelization. Future work will explore distilling the memory construction pipeline to further reduce latency for resource-constrained edge deployment. (ii) *Dependency on Upstream Perception:* As a modular sys-

tem, MM-MEM benefits from the rapid advancements in upstream vision encoders and captioners. While the system effectively filters irrelevant information via top-down retrieval, we anticipate that integrating stronger, end-to-end trained perception backbones will further enhance the system's robustness against visual artifacts. (iii) *Generalization to Unsupervised Scenarios:* The current memory manager utilizes task-driven reinforcement learning (SIB-GRPO) to align memory retention with reasoning needs. Extending this mechanism to fully unsupervised or self-supervised settings—where explicit task signals are absent—represents an exciting direction for enabling autonomous, "lifelong" learning agents. (iv) *Evolution towards Lifelong Agents:* Our evaluation focuses on standard long-video benchmarks. To better address true real-world deployment involving continuous, multi-session interactions with distribution shifts, we plan to extend MM-MEM to support dynamic memory updating and forgetting mechanisms better suited for open-ended, continuous agentic scenarios.

Ethics Statement

Our research advances the capability of multimodal agents to process and remember long-form video content. We recognize the importance of responsible AI development and address the following considerations. (i) *Privacy and Data Protection:* Memory-augmented agents inevitably process visual data that may contain personally identifiable information (PII). While our experiments utilize public, consented datasets, real-world deployment requires strict adherence to data minimization principles. We advocate for implementing local storage solutions and rigorous access controls to prevent unintended data disclosure. (ii) *Bias Mitigation:* Intelligent agents may inherit biases present in the training data or upstream foundation models. The selective nature of memory construction could theoretically retain biased evidence if not carefully managed. We encourage continuous monitoring of the memory selection policy and the adoption of diverse benchmarks to ensure fair and representative reasoning outcomes. (iii) *Responsible Deployment:* As with any powerful multimodal system, there is a potential for misuse in high-stakes decision-making. We emphasize that MM-MEM is designed as an assistive tool to augment human capabilities. Deployments in sensitive domains should always incorporate human-in-the-loop oversight to ensure reliability and accountability.

Use of AI Assistants

We used large language models (e.g., ChatGPT and Gemini) in a lawful and policy-compliant manner solely for non-substantive assistance such as translation and language polishing of the manuscript. They were not used to generate experimental results, derive scientific claims, or make methodological decisions; all technical content and conclusions are authored and verified by the authors.

Acknowledgments

We thank the anonymous reviewers and chairs for their efforts and constructive suggestions. This work is supported in part by the National Natural Science Foundation of China under grants 62521006, 624B2088, 62536003, 62571298, and 62576122.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Xinye Cao, Hongcan Guo, Jiawen Qian, Guoshun Nan, Chao Wang, Yuqi Pan, Tianhao Hou, Xiaojuan Wang, and Yutong Gao. 2025. Videominer: Iteratively grounding key frames of hour-long videos via tree-based group relative policy optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23773–23783.

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

Cody V Dong, Qihong Lu, Kenneth A Norman, and Sebastian Michelmann. 2025. Towards large language models with human-like episodic memory. *Trends in Cognitive Sciences*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2025a. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, et al. 2025b. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*.

Kuofeng Gao, Yang Bai, Jindong Gu, Shu-Tao Xia, Philip Torr, Zhifeng Li, and Wei Liu. 2024a. Inducing high energy-latency of large vision-language models with verbose images. In *ICLR*.

Kuofeng Gao, Huanqia Cai, Qingyao Shuai, Dihong Gong, and Zhifeng Li. 2024b. Embedding self-correction as an inherent ability in large language models for enhanced mathematical reasoning. *arXiv preprint arXiv:2410.10735*.

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.

Shaul Hochstein and Merav Ahissar. 2002. View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36(5):791–804.

Zhengjun Huang, Zhoujin Tian, Qintian Guo, Fangyuan Zhang, Yingli Zhou, Di Jiang, Zeying Xie, and Xiaofang Zhou. 2025. Licomemory: Lightweight and cognitive agentic memory for efficient long-term reasoning. *arXiv preprint arXiv:2511.01448*.

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

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer.

Niu Lian, Jun Li, Jinpeng Wang, Ruisheng Luo, Yaowei Wang, Shu-Tao Xia, and Bin Chen. 2025. Autossvh: Exploring automated frame sampling for efficient self-supervised video hashing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18881–18890.

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*, pages 5971–5984.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

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, et al. 2024. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *arXiv preprint arXiv:2411.13093*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. 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)*, pages 12585–12602.

Multi-Linguality Multi-Functionality Multi-Granularity. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. 2023. Memgpt: Towards llms as operating systems.

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, et al. 2025. Hd-epic: A highly-detailed egocentric video dataset. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23901–23913.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. 2024. Streaming long video understanding with large language models. *Advances in Neural Information Processing Systems*, 37:119336–119360.

Valerie F Reyna and Charles J Brainerd. 1995. Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences*, 7(1):1–75.

Xiaoqian Shen, Wenxuan Zhang, Jun Chen, and Mohamed Elhoseiny. 2025. Vgent: Graph-based retrieval-reasoning-augmented generation for long video understanding. *arXiv preprint arXiv:2510.14032*.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. 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*, pages 18221–18232.

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

Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Jinpeng Wang, Niu Lian, Jun Li, Yuting Wang, Yan Feng, Bin Chen, Yongbing Zhang, and Shu-Tao Xia. 2025a. Efficient self-supervised video hashing with selective state spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7753–7761.

Junke Wang, Dongdong Chen, Chong Luo, Xiyang Dai, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. 2023. Chatvideo: A tracklet-centric multimodal and versatile video understanding system. *arXiv preprint arXiv:2304.14407*.

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

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024b. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer.

Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang. 2024c. Longllava: Scaling multi-modal llms to 1000 images efficiently via a hybrid architecture. *arXiv preprint arXiv:2409.02889*.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2025b. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3272–3283.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. 2024. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Kristian Kersting, Jeff Z Pan, Hinrich Schütze, et al. 2025. Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. *arXiv preprint arXiv:2508.19828*.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. 2024a. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024b. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024c. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. 2025. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13691–13701.

A Proof of the Variational IB Bounds

We consider a stochastic encoder (memory manager) $p_\theta(m \mid x)$ that maps sensory input X to an episodic representation M. We assume the standard IB Markov structure $Y \leftrightarrow X \leftrightarrow M$[1], hence

[1]This follows the standard Information Bottleneck assumption (Tishby et al., 2000): M is extracted solely from X, i.e., M and Y are conditionally independent given X.

$$p_\theta(x, y, m) = p(x, y)\, p_\theta(m \mid x), \tag{14}$$
$$p_\theta(y \mid x, m) = p(y \mid x). \tag{15}$$

The induced marginals are

$$p_\theta(m) = \int p(x)\, p_\theta(m \mid x)\, dx, \tag{16}$$
$$p_\theta(m, y) = \int p(x, y)\, p_\theta(m \mid x)\, dx. \tag{17}$$

We introduce (i) a variational decoder $q_\phi(y \mid m)$ to approximate $p_\theta(y \mid m)$, and (ii) a variational prior $r(m)$ to approximate the intractable marginal $p_\theta(m)$. Define

$$\mathcal{L}_{\mathrm{p}}(\theta, \phi) \triangleq \mathbb{E}_{p(x,y)\, p_\theta(m|x)}[\log q_\phi(y \mid m)], \tag{18}$$

$$\mathcal{L}_{\mathrm{c}}(\theta) \triangleq \mathbb{E}_{p(x)}[D_{\mathrm{KL}}(p_\theta(m \mid x) \,\|\, r(m))]. \tag{19}$$

A.1 Lower bound on $I(M; Y)$

Step 1: rewrite mutual information.

$$\begin{aligned} & I(M; Y) \\ &= \mathbb{E}_{p_\theta(m,y)}\left[\log \frac{p_\theta(y \mid m)}{p(y)}\right] \\ &= \mathbb{E}_{p_\theta(m,y)}[\log p_\theta(y \mid m)] + H(Y). \end{aligned} \tag{20}$$

Step 2: variational lower bound via KL non-negativity. For each m, by the non-negativity of KL divergence,

$$\begin{aligned} & D_{\mathrm{KL}}(p_\theta(y \mid m) \,\|\, q_\phi(y \mid m)) \\ &= \mathbb{E}_{p_\theta(y|m)}[\log p_\theta(y \mid m) - \log q_\phi(y \mid m)] \\ &\geq 0. \end{aligned} \tag{21}$$

which implies

$$\begin{aligned} & \mathbb{E}_{p_\theta(m,y)}[\log p_\theta(y \mid m)] \\ &\geq \mathbb{E}_{p_\theta(m,y)}[\log q_\phi(y \mid m)]. \end{aligned} \tag{22}$$

Step 3: match the training sampling form. Using $p_\theta(m, y) = \int p(x, y)\, p_\theta(m \mid x)\, dx$, we have

$$\begin{aligned} & \mathbb{E}_{p_\theta(m,y)}[\log q_\phi(y \mid m)] \\ &= \mathbb{E}_{p(x,y)\, p_\theta(m|x)}[\log q_\phi(y \mid m)] \\ &= \mathcal{L}_{\mathrm{p}}(\theta, \phi). \end{aligned} \tag{23}$$

Combining with (20)–(22),

$$I(M; Y) \geq \mathcal{L}_{\mathrm{p}}(\theta, \phi) + H(Y). \tag{24}$$

A.2 Upper bound on $I(X; M)$

Step 1: rewrite mutual information.

$$\begin{aligned} & I(X; M) \\ &= \mathbb{E}_{p(x)}[D_{\mathrm{KL}}(p_\theta(m \mid x) \,\|\, p_\theta(m))]. \end{aligned} \tag{25}$$

Step 2: relate $\mathcal{L}_{\mathbf{c}}(\theta)$ and $I(X; M)$.

$$\begin{aligned} & \mathcal{L}_{\mathrm{c}}(\theta) \\ &= \mathbb{E}_{p(x,m)}\left[\log \frac{p_\theta(m \mid x)}{p_\theta(m)}\right] + \mathbb{E}_{p(x,m)}\left[\log \frac{p_\theta(m)}{r(m)}\right] \\ &= I(X; M) + \mathbb{E}_{p_\theta(m)}\left[\log \frac{p_\theta(m)}{r(m)}\right] \\ &= I(X; M) + D_{\mathrm{KL}}(p_\theta(m) \,\|\, r(m)) \\ &\geq I(X; M). \end{aligned} \tag{26}$$

Therefore,

$$I(X; M) \leq \mathcal{L}_{\mathrm{c}}(\theta). \tag{27}$$

A.3 Variational objective

Combining the above bounds,

$$\begin{aligned} & I(X; M) - \beta I(M; Y) \\ &\leq \mathcal{L}_{\mathrm{c}}(\theta) - \beta\Big(\mathcal{L}_{\mathrm{p}}(\theta, \phi) + H(Y)\Big). \end{aligned} \tag{28}$$

Since $H(Y)$ is constant w.r.t. (θ, ϕ), minimizing $I(X; M) - \beta I(M; Y)$ is equivalent to maximizing

$$\max_{\theta, \phi} \quad \beta\, \mathcal{L}_{\mathrm{p}}(\theta, \phi) - \mathcal{L}_{\mathrm{c}}(\theta). \tag{29}$$

This completes the proof.

Remark: Action-output correspondence. To avoid ambiguity, we clarify that the operator output $a_{t,i} \in \mathcal{O}$ uniquely determines the update outcome of the episodic stream. Thus, for a fixed sensory buffer $\mathcal{M}_{\mathrm{sens}}$ and fixed update rules, specifying the action sequence is sufficient to determine the resulting episodic memory.

Concretely, for each sensory item $m_{t,i}$ and the current latest node $e^\star$, the decision operator $\psi(\cdot)$ produces an action $a_{t,i}$, and the stream is updated by a deterministic transition

$$\begin{aligned} e^\star_{t,i} &= T\big(e^\star_{t,i-1},\, m_{t,i},\, a_{t,i}\big), \\ \mathcal{M}^{t,i}_{\mathrm{epi}} &= U\Big(\mathcal{M}^{t,i-1}_{\mathrm{epi}},\, m_{t,i},\, a_{t,i}\Big), \end{aligned} \tag{30}$$

where $T(\cdot)$ and $U(\cdot)$ are deterministic functions parameterized by the selected action $a_{t,i} \in$

$\{\text{ADD\_NEW}, \text{MERGE}, \text{DISCARD}\}$. Therefore, the final episodic stream is a deterministic function of the action sequence:

$$\mathcal{M}_{\text{epi}} = F(\mathcal{M}_{\text{sens}}, \{a_{t,i}\}) . \tag{31}$$

In other words, under fixed rules (T, U), the action space $\mathcal{O}$ and the resulting episodic memory $\mathcal{M}_{\text{epi}}$ provide equivalent descriptions of the same compression process.

B Dataset and Benchmark Details

B.1 Proposed HD-EPIC++

Motivation and split protocol. Our method requires SIB-GRPO fine-tuning, which in turn necessitates a dedicated training split. However, we find that datasets commonly used in prior long-video methods are either too short in duration or lack paired VQA annotations (Lian et al., 2025; Wang et al., 2025a). Therefore, to enable direct and reproducible evaluation, we build upon the HD-EPIC dataset (Perrett et al., 2025) and further adapt it with a fixed train/test partition, so that different methods can be compared under a consistent split without repeated re-partitioning. Concretely, we reorganize the original HD-EPIC videos into **105** training videos and **51** test videos, totaling **156** videos. The training split is used for SIB-GRPO fine-tuning, while all benchmark results are reported exclusively on the held-out test split.

Dataset overview. We introduce HD-EPIC++, an improved dataset built upon HD-EPIC, extending it with denser, fine-grained annotations tailored for long-horizon, procedure-centric video understanding. Compared to the original HD-EPIC, HD-EPIC++ emphasizes richer supervision signals that support not only recognition and temporal grounding, but also structured reasoning over procedures, entities, and interactions.

Dense annotations. HD-EPIC++ provides dense supervision covering **7** key annotation types: *(i) Recipe* (identify, retrieve, and localize recipes and steps), *(ii) Ingredient* (track ingredient usage, weight, timing, and order), *(iii) Nutrition* (analyze ingredient nutrition and its evolution throughout recipes), *(iv) Fine-Grained Action* (understand the what/how/why of actions), *(v) 3D Perception* (reason about object positions in 3D space), *(vi) Object Motion* (track object movements across long video sequences), and *(vii) Gaze* (estimate fixation points and anticipate future interactions).

VQA benchmark construction. Leveraging the dense annotations, we construct a multiple-choice VQA benchmark to evaluate long-context, multimodal reasoning. For each question type, we use a **5-way** multiple-choice format. We design **30** question prototypes, which instantiate into **26,650** multiple-choice questions. To increase difficulty and reduce shortcut learning, we sample **hard negatives** from within the dataset based on the underlying annotations.

Scalability and intended impact. The benchmark is designed to be large-scale yet tractable for evaluation with closed-source VLMs. Due to the annotation density, we estimate an upper bound of approximately **100,000** unique questions that can be generated via additional instantiations. We expect HD-EPIC++ to facilitate systematic evaluation of (i) long-horizon procedural understanding, (ii) entity/state tracking over time, and (iii) grounded multimodal reasoning under realistic distractors.

C Salient Key Sub-clip Extraction for Sensory Buffer

For each clip c_t with frames $\{f_{t,i}\}_{i=1}^{|c_t|}$, we quantify inter-frame variation by

$$d_{t,i} = \frac{1}{|\Omega|} \sum_{p \in \Omega} \left\| f_{t,i}(p) - f_{t,i-1}(p) \right\|_1 , \tag{32}$$

where Ω denotes the pixel grid. We then compute the clip-level statistics

$$\mu_t = \mathbb{E}_i[d_{t,i}], \qquad \sigma_t = \text{Std}_i[d_{t,i}], \tag{33}$$

and select salient indices via a simple adaptive threshold:

$$\mathcal{S}_t = \{\, i \mid d_{t,i} > \mu_t + \sigma_t \,\}. \tag{34}$$

Key sub-clip construction. For each $i \in \mathcal{S}_t$, we extract a short temporal window centered at i as a key sub-clip (sensory evidence), and record its temporal location $\tau_{t,i}$ (e.g., the timestamp of the center frame).

Near-duplicate suppression. To avoid redundant evidence while preserving salient dynamics, we suppress near-duplicate candidates among $\mathcal{S}_t$ by enforcing a minimum temporal separation. Concretely, let $\mathcal{S}_t$ be sorted by decreasing $d_{t,i}$, and iteratively keep an index i only if it is at least Δ frames away from all previously kept indices; otherwise it is discarded. This yields a compact set of

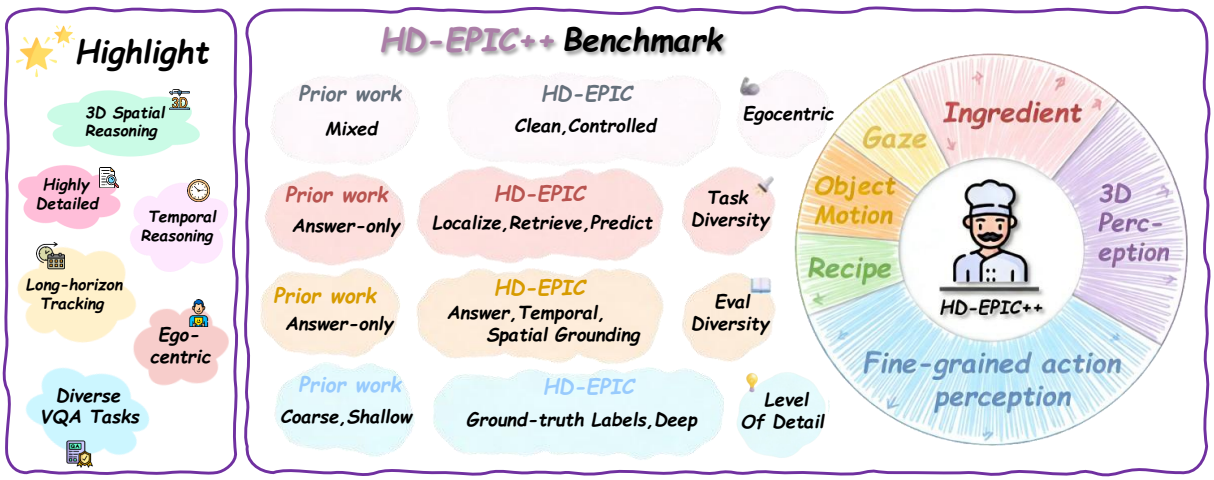


Figure 5: **HD-EPIC++** is an egocentric long-horizon kitchen video benchmark with highly detailed annotations, covering fine-grained action perception, temporal reasoning, 3D spatial understanding, object motion, gaze, and diverse VQA tasks (e.g., recipes and ingredients).

LoRA	
lora_rank	64
lora_alpha	128
lora_dropout	0.05
SIB-GRPO	
epoch	3
batch_size	8
learning_rate	1e-5
beta	0.1
ppo_clip_epsilon	0.2
use_importance_sampling	TRUE
kl_penalty_coef	0.1
save_steps	100
Reasoning	
top_k_sym	5
top_k_epi	2
top_k_sen	1
γ	0.72

Table 5: Training and inference hyperparameters.

key sub-clips that covers salient changes without excessive overlap.

Memory tuple instantiation. Each key sub-clip centered at i is encoded into a visual representation $\mathbf{v}_{t,i}$ (e.g., by a video encoder), paired with a text trace $\mathbf{l}_{t,i}$ (aligned subtitles or automatic captioning), and stored with its temporal location $\tau_{t,i}$.

D Implementation Details

Hyperparameter Settings. We fine-tune the model with SIB-GRPO on the training split of HD-EPIC++, and perform LoRA adaptation within the SWIFT framework. The hyperparameters used for training and inference are summarized in Table 5.

ID	Method	Video-MME			
		Short	Medium	Long	Overall
(I)	w/o $SIB-GRPO$	80.1	73.9	71.9	75.3
(II)	w/o $VisualMemory$	76.5	70.1	66.1	70.9
(III)	w/o $TextMemory$	80.5	74.2	72.8	75.8
(IV)	w/o $Symbolic$	81.2	74.6	73.5	76.4
(V)	w/o $Episodic$	80.2	73.6	71.2	75.0
(VI)	w/o $Sensor$	78.6	71.9	70.2	73.6
(VII)	w/o $Memory$	79.2	72.6	70.9	74.2
(VIII)	**MM-MEM (full)**	**82.8** $\pm$ **0.2**	**75.8** $\pm$ **0.2**	**75.7** $\pm$ **0.3**	**78.1** $\pm$ **0.2**

Table 6: Ablation studies of individual modules on Video-MME with subtitles.

Evaluation protocol. The same evaluation protocol as Flash-VStream (Zhang et al., 2024a) is followed. Since VStream-QA consists of open-ended questions, GPT-4o-mini (Hurst et al., 2024) is adopted as an automatic judge. Given a model prediction, whether the prediction is correct is judged, and a score between 0 and 5 is assigned. Accordingly, two metrics are reported on VStream-QA: (i) Accuracy, which indicates whether the prediction is judged correct, and (ii) Score, which is computed as the average of the assigned scores.

E Additional Experiments

E.1 Ablation Studies

Table 6 reports module-wise ablations on Video-MME under the subtitle setting. The full model (MM-MEM) achieves **82.8/75.8/75.7** on Short-/Medium/Long and **78.1** overall. Removing any component consistently degrades performance, indicating that each module contributes positively.

Method	VS-Ego		VS-Movie	
	Acc.	Sco.	Acc.	Sco.
Video-ChatGPT	51.7	3.7	54.4	3.4
MovieChat	52.2	3.4	39.1	2.3
Chat-UniVi	50.9	3.8	54	3.4
LLaMA-VID	54.8	3.9	51.4	3.4
Flash-VStream	59	3.9	56.1	3.4
MM-MEM	62.5	4.1	52.1	3.2

Table 7: Performance comparison on the VStream-QA benchmark, evaluated on the VS-Ego and VS-Movie splits in terms of accuracy (Acc.) and score (Sco.).

Visual vs. Text Memory. Visual Memory is the most critical component: removing it drops Overall from 78.1 to 70.9, with an even larger decrease on Long (75.7 $\rightarrow$ 66.1, **-9.6**). In contrast, removing Text Memory yields a smaller but consistent drop (Overall 78.1 $\rightarrow$ 75.8, **-2.3**), suggesting that subtitles provide useful high-level cues, yet fine-grained and verifiable visual evidence remains indispensable, especially for long videos.

Training/Memory Management. Removing SIB-GRPO reduces Overall to 75.3 (**-2.8**) and Long to 71.9 (**-3.8**), showing that the redundancy-aware memory optimization is particularly beneficial as temporal context grows.

Hierarchical Memory Components. Among the three memory layers, *Sensor* and *Episodic* have stronger impact on longer videos: removing Sensor yields Overall 73.6 (**-4.5**) and Long 70.2 (**-5.5**), while removing Episodic gives Overall 75.0 (**-3.1**) and Long 71.2 (**-4.5**). Removing Symbolic causes a smaller average drop (Overall 76.4, **-1.7**), but remains consistently helpful, likely supporting higher-level temporal/relational reasoning.

Disabling memory entirely. Without the memory mechanism, performance drops to Overall 74.2 (**-3.9**) and Long 70.9 (**-4.8**), indicating that explicit memory retrieval/organization improves robustness even when subtitles are available.

E.2 Descriptive Statistics

For the main ablation in Table 6, we report mean accuracy with uncertainty estimated from repeated evaluations, shown as $\pm$ values. The full MM-MEM achieves $\mathbf{82.8 \pm 0.2}$, $\mathbf{75.8 \pm 0.2}$, and $\mathbf{75.7 \pm 0.3}$ accuracy on the Short/Medium/Long subsets, respectively, with an overall score of $\mathbf{78.1 \pm 0.2}$. The relatively small variances across subsets indicate stable performance under the same evaluation protocol and suggest that the improvements of the full system over ablated variants are consistent rather than driven by outlier runs.

E.3 Results on VSTREAM-QA

Table 7 summarizes results on the VStream-QA benchmark, covering two splits: VS-Ego and VS-Movie. Overall, MM-MEM achieves the best performance on the VS-Ego split, while its performance on VS-Movie is comparatively weaker, suggesting different challenges across domains.

VS-Ego (streaming egocentric videos). MM-MEM attains the highest **Acc.** of **62.5** and the highest **Sco.** of **4.1**. Compared with the strongest baseline Flash-VStream (59.0 Acc., 3.9 Sco.), MM-MEM improves accuracy by **+3.5** and score by **+0.2**. This indicates that MM-MEM is particularly effective for streaming egocentric understanding, where long-horizon context accumulation and memory utilization are crucial.

VS-Movie (streaming movie clips). On VS-Movie, MM-MEM achieves 52.1 Acc. and 3.2 Sco., remaining below the best-performing methods (e.g., Video-ChatGPT, Chat-UniVi, and Flash-VStream, which achieve $\sim$54.0–56.1 Acc. and 3.4 Sco.). In particular, compared with Flash-VStream, MM-MEM lags by **4.0** points in accuracy (52.1 vs. 56.1) and **0.2** points in score (3.2 vs. 3.4). This suggests that streaming QA in the movie domain may rely on cues such as cinematographic scene transitions, dialogue style, and narrative coherence, which are not yet fully captured by the current memory formulation.

Cross-split observation. While MM-MEM yields clear gains on VS-Ego, its relative advantage diminishes on VS-Movie, implying that the proposed memory-centric design generalizes better to egocentric streaming scenarios than to movie-style content under the current setup.

E.4 SFT vs. SIB-GRPO RL

As shown in Table 8, supervised fine-tuning (SFT) yields a limited improvement over the Qwen3-VL-8B baseline (from 25.88 to 27.25, +1.37). Conversely, SIB-GRPO achieves 30.28, providing a substantial +4.40 gain over the baseline and +3.03 over SFT. While SFT primarily aligns the model with surface-level training distributions, SIB-GRPO directly optimizes decision-making for long-horizon reasoning. This demonstrates that

Method	HD-EPIC++ Accuracy
Qwen3-VL-8B (Bai et al., 2025)	25.88
Qwen2.5-VL-7B (Bai et al., 2025)	24.37
LLaVA-Video-7B (Bai et al., 2025)	25.37
VideoLLaMA 3-7B (Zhang et al., 2025)	20.36
Qwen3-VL-4B (Bai et al., 2025)	24.91
Qwen3-VL-2B (Bai et al., 2025)	22.80
Qwen3-VL-8B (sft)	27.25
MM-MEM (Ours)	**30.28**

Table 8: Evaluation on the proposed **HD-EPIC++**.

preference-driven RL effectively mitigates compounding misinterpretations—the primary error source in HD-EPIC++ by rewarding correct end-task behavior rather than merely enhancing visual feature extraction.

E.5 Memory Compression Operations

To illustrate our IB-based compression, we analyze a long "cooking" video from **HD-EPIC++**, demonstrating how MM-MEM transitions sensory memory into episodic memory. ADD_NEW captures semantically novel shifts (e.g., from "chopping ingredients" to "heating the pan"). MERGE consolidates continuous actions (e.g., consecutive segments of "chopping an onion"). DISCARD eliminates noisy or redundant frames (e.g., motion blur from camera shake). Overall, MM-MEM achieves semantically guided compression, efficiently retaining salient events while filtering out redundancy.

E.6 Robustness to PySceneDetect Errors

PySceneDetect relies on basic heuristics, occasionally causing coarse over- or under-segmentation. However, these errors are non-critical as MM-MEM uses it solely for initialization. Similar to (Gao et al., 2024b), our memory formation exhibits inherent self-correction: the Episodic Stream resolves over-segmentation via MERGE and DISCARD actions, while adaptive sampling mitigates under-segmentation. Ultimately, MM-MEM's robustness stems from the synergy of adaptive sampling and RL-based memory consolidation, rather than the initial segmentation's precision.

F Details of Prompts

Prompt sources. The Answer Agent prompt in Figure 6 follows the template described in the official Qwen3-VL technical report (Bai et al., 2025), which aligns with our choice of Qwen3-VL as the base model. Similarly, the evaluation agent prompts for VStream-QA (Zhang et al., 2024a) (system and user prompts in Figure 7a and Figure 7b) are adopted from the prompts provided on the official VStream-QA website.

Answer Agent

Answer Prompt

```
"Select the best answer to the following multiple-choice question based on the video.\n"
"Respond with only the letter (A, B, C, or D) of the correct option.\n"
f"Question: {question_text} Possible answer choices:\n"
f"{options_str}\n"
"The best answer is:"
```

Figure 6: Answer Agent prompt template instructing the model to select the best option for a video-based multiple-choice question and respond only with the corresponding letter (A–D).

Evaluation Agent

System Prompt

```
"You are an intelligent chatbot designed for evaluating the correctness of generative outputs for question-answer pairs. "
"Your task is to compare the predicted answer with the correct answer and determine if they match meaningfully. Here's how you can accomplish the task:"
"- Focus on the meaningful match between the predicted answer and the correct answer.\n"
"- Consider synonyms or paraphrases as valid matches.\n"
"- Evaluate the correctness of the prediction compared to the answer."
```

(a) System prompt for VStream-QA evaluation agent, instructing the model to judge whether a generated answer matches the ground-truth answer.

Evaluation Agent

User Prompt

```
"Please evaluate the following video-based question-answer pair:\n\n"
f"Question: {question}\n"
f"Correct Answer: {answer}\n"
f"Predicted Answer: {pred}\n\n"
"Provide your evaluation only as a yes/no and score where the score is an integer value between 0 and 5, with 5 indicating the highest meaningful match. "
"Please generate the response in the form of a Python dictionary string with keys 'pred' and 'score', where value of 'pred' is a string of 'yes' or 'no' and value of 'score' is in INTEGER, not STRING."
"DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. "
"For example, your response should look like this: {'pred': 'yes', 'score': 4}."
```

(b) User prompt for the VStream-QA evaluation agent, requesting a strict Python-dictionary output.

Figure 7: Prompts used by the VStream-QA evaluation agent (system and user prompts).