AEMAgent: Cognition-Inspired Active Memory Formation for Very-Long Video Understanding

Anonymous ACL submission

Abstract

Embodied AI and wearable assistants require agents to retrieve sparse evidence from continuous days-level videos. Existing memory-based methods typically follow a passive recording paradigm based on fixed-window captioning and periodic memory construction, which produces redundant memories and inefficient long-horizon reasoning. Inspired by human memory formation, we propose Active Explore Memory Agent (**AEMAgent**), which transforms passive recording into active memory formation by deciding *when to remember*, *how to explore events*, and *when to reflect*. These decisions are instantiated as a coherent pipeline: **Surprise-guided Event Selection** detects salient event changes for selective memory writing, **Exploratory Episodic Memory Encoding** enriches accepted events through self-questioning, **Predictive-Coding-Guided Reflection** updates knowledge from expectation-observation mismatches, and **Semantic Memory Consolidation** distills noisy episodic updates into compact long-term memory. Experiments on multiple benchmarks show that AEMAgent achieves the best overall performance among all compared methods while maintaining lower storage overhead than strong full-memory baselines, demonstrating the effectiveness of selective and efficient memory formation for scalable very-long video understanding.[1]

1 Introduction

Recent advances in embodied AI and wearable assistants such as AI glasses are extending video understanding from short curated clips to continuous, days-level streams (Yang et al., 2025; Park et al., 2026). In such scenarios, agents must answer questions about events that may have occurred days earlier, requiring them to retrieve sparse query-relevant evidence from long, repetitive, and weakly structured histories. This setting, known as very-long video understanding, poses challenges beyond conventional long-video understanding. Existing long-video methods, including context compression (Tang et al., 2025) and KV-cache compression (Kim et al., 2026), remain fundamentally constrained by fixed-size context windows and cannot preserve persistent historical information over extended time spans. As a result, recent work has increasingly explored long-term memory mechanisms for very-long video understanding (Long et al., 2025; Yeo et al., 2025).

Although recent agent-based methods introduce long-term memory, most still follow a static observation paradigm. As shown on the left side of Figure 1, they typically segment videos into fixed temporal windows, generate caption-only records, and periodically aggregate them into semantic memory. This process leads to three key limitations. First, rigid temporal slicing ignores the irregular rhythms of real-world events and forces agents to process predefined intervals rather than meaningful changes. Second, caption-centered memory construction mainly captures surface-level visual content, without exploring deeper event implications. Third, periodic aggregation updates semantic memory regardless of actual information gain, introducing redundant content into long-term memory. Consequently, such static memory formation often produces redundant, shallow, and limited-coverage memories, increasing retrieval burden and contextual interference during question answering.

A more capable very-long video agent should move beyond static observation by actively deciding *when to remember*, *how to explore events*, and *when to reflect*. This desired capability is naturally aligned with human memory formation (Atkinson and Shiffrin, 1968). Humans do not uniformly record continuous experience, instead, they selectively encode salient changes in the environment. Humans also actively explore events by asking

[1] Our code will be released in the camera-ready version.

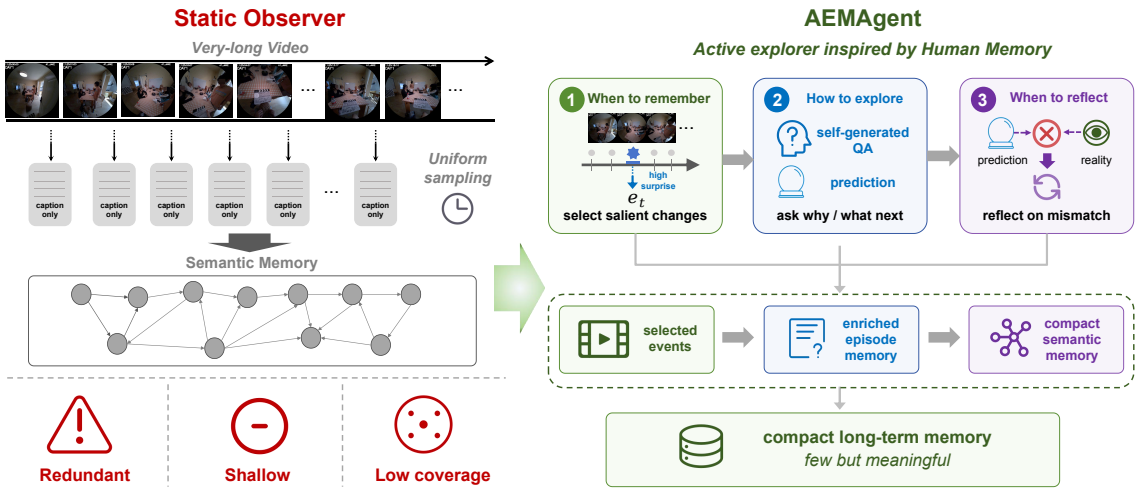


Figure 1: **From static observer to cognition-inspired active explorer.** Static observation builds memory from fixed-interval caption-only samples, often yielding redundant, shallow, and low-coverage memories. In contrast, AEMAgent transforms passive observation into cognition-inspired active exploration, producing fewer but more meaningful memories for very-long video understanding.

implicit questions about their causes, meanings, and possible consequences (Zacks et al., 2007). Moreover, cognitive theories such as predictive coding (Rao and Ballard, 1999) suggest that mismatches between prior expectations and actual observations can trigger belief revision. These principles provide a natural blueprint for building an active memory agent that selectively acquires, enriches, and consolidates long-term memory.

Inspired by these observations, we propose the Active Explore Memory Agent (**AEMAgent**), a cognition-inspired memory agent for very-long video understanding. AEMAgent transforms long-term memory construction from passive recording into active memory formation by addressing three key questions: *when to remember*, *how to explore events*, and *when to reflect*. As illustrated in the right part of Figure 1, AEMAgent first decides *when to remember* by adaptively detecting salient event transitions from visual and audio observations. It then determines *how to explore events* by enriching selected events through self-questioning and short-term prediction, producing episodic knowledge beyond surface-level captions. Finally, it decides *when to reflect* by using prediction–observation mismatches to trigger semantic memory updates, which are further consolidated into a compact and retrievable long-term semantic graph.

Extensive experiments on multiple very-long video understanding benchmarks demonstrate that AEMAgent achieves strong overall performance with a favorable efficiency–accuracy trade-off. By forming compact and query-relevant memories, AEMAgent reduces storage and retrieval overhead among full-memory baselines. These results suggest that scalable very-long video understanding requires not merely larger memory banks, but selective, exploratory, and efficient memory formation.

Our contributions are summarized as follows:

- We identify the limitation of existing very-long video memory agents as a static observation paradigm, where fixed-window captioning and periodic memory aggregation lead to redundant, shallow, and low-coverage memories.
- We propose AEMAgent, a cognition-inspired memory agent that replaces passive long-term memory recording with active memory formation. Through surprise-guided event selection, exploratory episodic encoding, and predictive-coding-guided reflection, AEMAgent autonomously decides *when to remember*, *how to explore events*, and *when to reflect*.
- Extensive experiments on multiple very-long video understanding benchmarks demonstrate that AEMAgent achieves the best overall performance while maintaining a strong efficiency–accuracy trade-off, highlighting the superiority of selective active memory formation over quantity-driven accumulation.

2 Related Work

2.1 Long Video Understanding

Long video understanding aims to enable models to comprehend videos spanning hours and answer questions that require temporally distributed evidence. Existing methods have made progress through extended context windows (Team et al., 2024; Ding et al., 2023), visual token compression (Huang et al., 2025; Li et al., 2024), sparse key-frame sampling (Tang et al., 2025; Song et al., 2026), hierarchical summarization (Song et al., 2024; Qian et al., 2024; Kim et al., 2026), and agent-based segment reasoning (Wang et al., 2024; Zhi et al., 2025). However, these approaches are primarily designed for minutes- or hours-level videos with bounded temporal scope. They either remain limited by finite context capacity, risk discarding sparse but critical evidence, or maintain only short-term segment-level memory within a single video. Therefore, they cannot be directly scaled to very-long video understanding, where agents must continuously retain, update, and retrieve information across days-level video histories.

2.2 Very-Long Video Understanding with Long-Term Memory

Very-long video understanding requires models to answer questions over continuous, repetitive, and weakly structured streams spanning hours to days. Recent agent-based methods address this challenge by constructing long-term memory from video streams and retrieving query-relevant evidence for response generation. EgoRAG (Yang et al., 2025) retrieves historical events from captions over fixed 30-second windows, while M3-Agent (Long et al., 2025) and EGAgent (Rege et al., 2026) organize segment-level observations into entity-centric or scene-graph memories. WorldMM (Yeo et al., 2025) extends this paradigm to multiple temporal scales, building episodic, semantic, and visual memories from 3-minute, 10-minute, and 1-hour captions. Despite different memory structures, these methods largely rely on fixed-window captioning for memory construction, which can introduce redundant and low-coverage memories. In contrast, AEMAgent transforms static memory construction into cognition-inspired active memory exploration by selectively remembering, exploring, and reflecting on meaningful event changes.

2.3 Human Memory Formation

Human memory is not a uniform recording of continuous experience. Cognitive studies suggest that memory formation is selective and event-centered: humans preferentially encode salient environmental changes, unexpected transitions, and behaviorally relevant information (Atkinson and Shiffrin, 1968; Zacks et al., 2007). Beyond encoding, humans actively interpret events by understanding their causes, relations, and implications. These event-level experiences correspond to episodic memory, which preserves concrete observations about what happened, when it happened, and in what context. Through consolidation, repeated experiences and local observations are abstracted into semantic memory, which stores stable knowledge such as habits, preferences, social relations, and high-level regularities (Tulving et al., 1972). Predictive coding further explains how semantic memory is revised over time. Prior experience generates expectations about upcoming observations, while mismatches between expectation and reality signal belief revision (Rao and Ballard, 1999; Friston, 2005). These principles motivate AEMAgent to model very-long video streams as selective, reflective, and continuously updated memory formation.

3 Method

3.1 Overview

We propose AEMAgent, a cognition-inspired agent that converts fixed-window passive recording into active memory formation for very-long video understanding. This process is organized around three decisions: *when to remember*, *how to explore events*, and *when to reflect*.

As illustrated in Figure 2, AEMAgent first applies *Surprise-Guided Event Selection* to detect coherent and salient event clips from continuous video streams. For each selected event, *Exploratory Episodic Memory Encoding* stores a representative frame as visual memory and writes captions and self-generated question-answer pairs into episodic memory, while generating a short-term prediction for future reflection. *Predictive-Coding-Guided Reflection* then compares predictions with subsequent observations and updates semantic memory only when informative mismatches occur. These semantic updates are further organized by *Semantic Memory Consolidation* into a compact and retrievable long-term semantic graph. During eval-

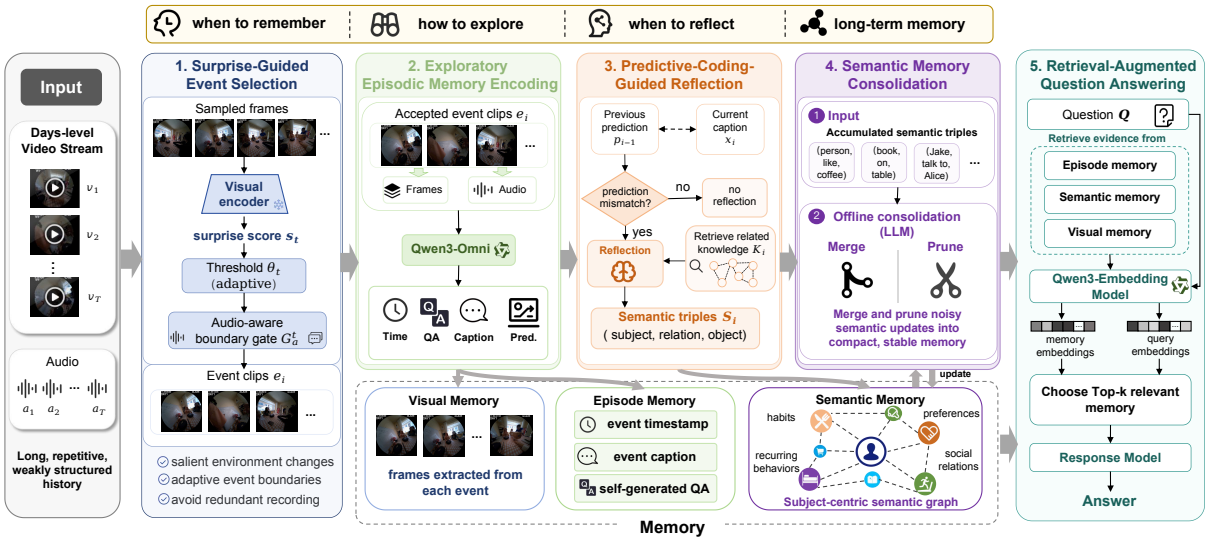


Figure 2: **Framework of AEMAgent.** AEMAgent actively forms visual, episodic, and semantic memories for very-long video understanding. It selects salient event changes, encodes them into exploratory episodic memories, updates semantic knowledge from prediction–observation mismatches, and consolidates compact long-term memory for retrieval-augmented question answering.

Algorithm 1 Surprise-Guided Event Selection

Require: Video stream V, encoder F, thresholds θ_v, ρ, update rate α, minimum duration $\Delta_{\min}$
Ensure: Event set $\mathcal{E}$
Initialize $\mathcal{E} \leftarrow \emptyset$, surprise history $\mathcal{H} \leftarrow \emptyset$, background $\mathbf{b}_0$, and $t_{\text{last}} \leftarrow 0$
for each sampled frame v_t at timestamp t **do**
  $\mathbf{z}_t \leftarrow \text{Norm}(F(v_t)), \quad s_t^v \leftarrow 1 - \mathbf{z}_t^\top \mathbf{b}_{t-1}$
  $\bar{s}_t \leftarrow \text{Mean}(\mathcal{H}), \quad C_t^v \leftarrow (s_t^v > \theta_v) \wedge (s_t^v > \rho \bar{s}_t)$
  $G_t^a \leftarrow \text{AudioGate}(V, t)$ using VAD
  if $C_t^v \wedge G_t^a \wedge (t - t_{\text{last}} > \Delta_{\min})$ **then**
    $\mathcal{E} \leftarrow \mathcal{E} \cup \{\text{Extract}(V, t_{\text{last}}, t)\}$
    $t_{\text{last}} \leftarrow t$
  end if
  $\mathbf{b}_t \leftarrow \text{Norm}(\alpha \mathbf{z}_t + (1-\alpha)\mathbf{b}_{t-1})$, update $\mathcal{H}$ with s_t^v
end for
return $\mathcal{E}$

uation, *Retrieval-Augmented Question Answering* retrieves relevant episodic, semantic, and visual memories as grounded evidence for the response model. Through this pipeline, AEMAgent enables selective memory writing, exploratory event understanding, and mismatch-driven knowledge updating for scalable very-long video reasoning.

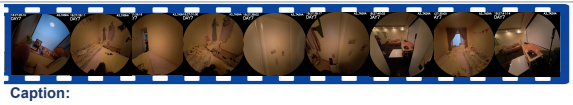
3.2 Surprise-Guided Event Selection

Very-long videos contain substantial routine content, making uniform frame processing inefficient. Inspired by human selective event segmentation, AEMAgent uses a streaming surprise detector to identify salient event changes for memory writing. Visual surprise serves as the primary cue for environmental changes, while audio continuity acts as an auxiliary gate to avoid interrupting ongoing speech or continuous background sounds.

The procedure is summarized in Algorithm 1. For each sampled frame v_t at timestamp t, a frozen DINOv3 (Siméoni et al., 2025) encoder F extracts a normalized representation $\mathbf{z}_t$. AEMAgent maintains an exponential moving average background state $\mathbf{b}_t$ to summarize recent visual context, and computes the visual surprise score s_t^v as the cosine distance between the current frame and the previous background state. Given the recent surprise history $\mathcal{H}$ and its local average $\bar{s}_t$, a frame is marked as visually salient when s_t^v exceeds both the absolute threshold θ_v and the relative threshold $\rho\bar{s}_t$. A candidate boundary is accepted only if it also passes an audio-aware gate G_t^a, estimated from voice activity detection (VAD) (Sohn et al., 1999), and is at least $\Delta_{\min}$ away from the previous boundary t_{last}. Once accepted, the segment between t_{last} and t is extracted as an event clip with its original audio track and metadata.

3.3 Exploratory Episodic Memory Encoding

After selecting an event clip, AEMAgent converts it into an exploratory memory unit using Qwen3-Omni-30B-A3B (Xu et al., 2025), an omni-modal model that jointly processes the event video and audio track to capture visual observations and acoustic cues. For each event, AEMAgent also stores a representative frame as visual memory, enabling



Caption:
[visual] Tasha moves through a dimly lit bedroom, passing a bed with a yellow comforter and a window with a view.....
[audio] A cat meows
[overall caption] Tasha walks from a bedroom into a room with a projector screen, with a cat meowing in the background.

Self-generated QA:
Q1: Who is the person moving through the rooms?
A1: Based on the timestamp, the person is identified as Tasha.
Q2: What is the purpose of the room with the projector screen?
A2: The room with the projector screen is likely used for presentations or watching movies.
...

Prediction:
Tasha will walk into the room with the projector screen and approach the table.

Figure 3: Example of exploratory episodic memory encoding. AEMAgent converts a selected event clip into a factual caption, self-generated QA pairs, and a next-event prediction.

later retrieval of fine-grained visual details.

As illustrated in Figure 3, the exploratory episodic record contains three components: a compact first-person caption summarizing visual and audio evidence, self-generated question-answer pairs that explore event participants, scene functions, and potential significance, and a short-term prediction of what may happen next. The prompt templates are provided in Appendix A.1. The caption and QA pairs are written into episodic memory, while the prediction is retained as an expectation signal for subsequent predictive-coding-guided reflection. In this way, AEMAgent turns each selected event from a shallow caption record into an exploratory memory unit that supports both retrieval and later reflection.

3.4 Predictive-Coding-Guided Reflection

Episodic memories provide concrete but local event-level evidence. Inspired by predictive coding, where expectation violations drive belief revision, AEMAgent introduces *Predictive-Coding-Guided Reflection* to update semantic memory from informative event transitions.

We represent semantic memory as a subject-centric directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{R})$, where nodes denote people, objects, concepts, or behaviors, and edges encode typed semantic relations. For consecutive event memories m_{i-1} and m_i, an LLM-based prediction-error judge compares the previous prediction p_{i-1} with the current factual caption x_i and determines whether the observation violates the expectation. The prompt template for prediction-error judging is provided in Appendix A.2. If no prediction error is detected, AEMAgent skips reflection and leaves $\mathcal{G}$ unchanged.

When a prediction error occurs, AEMAgent retrieves relevant semantic context K_i from $\mathcal{G}$ and invokes an LLM to reflect over x_{i-1}, p_{i-1}, x_i, the mismatch explanation, and K_i. The reflection module generates candidate semantic triples S_i that capture causal relations, state transitions, and expectation violations. The prompt template is provided in Appendix A.3. A local grounding check then filters unsupported triples, and only validated triples $\tilde{S}_i$ are inserted into $\mathcal{G}$. If an equivalent edge already exists, its timestamp is refreshed as lightweight memory reinforcement. This step focuses on validating semantic updates for the current prediction error, leaving cross-event pruning and abstraction to the global consolidation stage.

3.5 Semantic Memory Consolidation

While predictive-coding-guided reflection updates semantic memory from local event transitions, the accumulated graph may become redundant, fragmented, or partially inconsistent over time. To maintain compact long-term memory, AEMAgent performs *Semantic Memory Consolidation* as an offline global organization step.

At the end of each day, AEMAgent applies LLM-based consolidation to the semantic graph. The prompt template for semantic memory consolidation is provided in Appendix A.4. Unlike local grounding in predictive reflection, consolidation operates across semantic edges collected over multiple events. It merges equivalent relations, removes redundant or conflicting edges, and abstracts repeated event-level observations into stable high-level knowledge, such as habits, preferences, social relations, and routines. Thus, AEMAgent decouples online reflection from offline consolidation to organize semantic memory into compact, retrievable long-term memory for question answering.

3.6 Retrieval-Augmented Question Answering

At evaluation time, AEMAgent answers each question by retrieving evidence from episodic memory and semantic memory, while visual memory is accessed through temporal alignment. Given a question Q with answer choices $\mathcal{Y}$, we construct a query text from the question and choices, and perform dense retrieval over textual memory entries. Specifically, the query and memory entries are encoded into normalized embeddings via Qwen3-Embedding-4B (Zhang et al., 2025), and

Table 1: **Task-wise QA performance on EgoLifeQA and Ego-R1 Bench.** Results are reported as Mean@5 accuracy (%) over five repeated runs. The Avg. reports the sample-size-weighted average across task categories. For each response model, we compare random retrieval and five agent-based memory methods. EL, ER, HI, RM, TM, and FP denote EntityLog, EventRecall, HabitInsight, RelationMap, TaskMaster, and Future Plan, respectively. **Bold** and underlined scores indicate the best and second-best results within each response-model block.

Response Model	Method	EgoLifeQA (%)						Ego-R1 Bench (%)						
		EL	ER	HI	RM	TM	Avg.	EL	ER	HI	RM	TM	FP	Avg.
Qwen3.5 35B-A3B	RANDOM	28.0	31.6	31.8	**38.7**	35.2	33.0	30.2	33.2	42.0	38.0	41.6	33.9	39.8
	EGORAG	9.6	16.7	38.4	35.4	30.2	23.9	20.2	14.6	24.9	27.9	17.4	35.7	20.3
	M3-AGENT	20.2	30.0	27.5	30.2	36.8	28.2	20.3	32.0	31.9	33.7	38.6	36.5	30.2
	WORLDMM	29.1	34.4	33.4	27.0	38.7	31.7	**56.7**	**56.8**	41.4	42.6	46.5	50.7	52.2
	EGAGENT	37.6	40.9	42.6	34.8	54.8	40.5	36.5	38.2	46.1	34.7	**51.5**	52.5	39.2
	AEMAGENT(OURS)	**39.4**	**44.0**	**45.6**	33.4	**60.0**	**42.4**	48.6	56.3	**52.0**	**58.2**	45.7	**56.0**	**54.1**
Qwen3-VL 30B	RANDOM	30.6	33.2	36.7	28.5	40.3	32.7	31.0	33.0	37.1	30.3	45.7	38.2	36.1
	EGORAG	9.6	14.3	41.0	33.6	33.3	26.6	23.2	14.5	27.0	30.3	22.2	37.7	21.9
	M3-AGENT	26.8	33.5	30.3	26.4	40.1	30.5	30.7	35.0	31.9	29.7	35.1	33.7	32.6
	WORLDMM	27.0	38.3	42.0	34.1	42.2	35.4	44.7	39.6	38.7	34.0	43.8	42.0	40.0
	EGAGENT	31.4	36.1	46.7	33.7	**48.7**	37.2	34.2	36.6	37.2	35.7	44.4	37.0	37.7
	AEMAGENT(OURS)	**32.4**	**38.9**	**50.8**	**36.0**	44.4	**40.5**	**54.8**	**46.8**	**47.4**	**47.3**	**47.5**	**43.5**	**48.5**
Mistral-Small 3.1-24B	RANDOM	26.1	30.8	34.8	36.3	32.7	31.7	26.9	29.8	34.7	30.7	30.9	12.2	16.7
	EGORAG	11.2	15.9	34.4	35.2	34.9	24.2	20.5	16.8	26.8	30.7	19.2	**34.6**	22.0
	M3-AGENT	30.0	30.3	33.4	30.8	32.7	31.0	30.0	34.8	34.1	27.9	31.9	19.0	21.2
	WORLDMM	33.8	**46.2**	**48.9**	38.6	38.4	40.5	31.7	30.1	36.0	27.2	40.6	34.4	31.1
	EGAGENT	38.5	39.4	34.8	29.2	40.0	36.2	36.3	40.5	36.8	31.2	42.6	28.3	33.1
	AEMAGENT(OURS)	**39.2**	44.3	46.7	**40.5**	**46.0**	**42.6**	**45.8**	**44.9**	**43.6**	**37.3**	**43.2**	34.0	**43.0**
Gemma-3 27B	RANDOM	24.8	26.2	29.5	24.0	39.4	27.4	29.2	36.7	36.2	34.3	36.0	32.1	34.3
	EGORAG	10.2	15.1	**41.0**	27.2	28.6	21.8	23.6	17.6	30.2	30.2	18.4	31.5	22.9
	M3-AGENT	24.0	34.4	21.0	22.6	35.2	27.3	41.4	42.4	29.1	38.2	33.3	36.4	39.5
	WORLDMM	29.9	36.5	32.5	28.5	**44.4**	33.4	54.0	54.6	36.5	40.9	**48.5**	44.9	49.6
	EGAGENT	20.9	30.7	23.3	**41.3**	32.8	28.1	24.8	33.3	24.0	41.8	36.5	40.8	32.6
	AEMAGENT(OURS)	**31.2**	**40.5**	39.3	31.8	41.3	**36.0**	**59.0**	**56.8**	**45.3**	**48.2**	44.4	**48.2**	**53.8**

dot-product similarity is used for ranking. For each textual memory type $m \in \{\text{epi}, \text{sem}\}$, AEMAgent retrieves the top-K_m relevant memories $\mathcal{R}_Q^m$.

For visual evidence, AEMAgent does not retrieve frames independently. Instead, it uses the timestamps of the retrieved episodic memories to locate temporally aligned representative frames from visual memory, denoted as $\mathcal{R}_Q^{\text{vis}}$. The final evidence set $\mathcal{R}_Q = \{\mathcal{R}_Q^{\text{epi}}, \mathcal{R}_Q^{\text{sem}}, \mathcal{R}_Q^{\text{vis}}\}$ is provided to the response model together with the question and answer choices to generate the final prediction. Other retrieval strategies and the effect of K_m are analyzed in the experiments.

Table 2: Memory efficiency comparison. We report memory type, memory-token cost (Tok.), retrieval time per question (Time/Q), and average QA accuracy (Avg.) on Ego-R1 Bench. The response model is Qwen3-VL-30B. Epi. and Sem. denote episodic and semantic memory, respectively.

Method	**Epi.**	**Sem.**	**Tok. (K)** ↓	**Time/Q (s)** ↓	**Avg. (%)** ↑
EgoRAG	✓	–	**1,743**	**0.003**	21.9
M3Agent	✓	✓	14,194	1.415	32.6
WorldMM	✓	✓	12,941	6.732	40.0
EGAgent	✓	✓	53,333	7.495	37.7
AEMAgent	✓	✓	8,777	0.269	**48.5**

4 Experiments

Baselines and evaluation setup. We compare AEMAgent with five baselines: Random Sample, EgoRAG (Yang et al., 2025), M3-Agent (Long et al., 2025), WorldMM (Yeo et al., 2025), and EGAgent (Rege et al., 2026). Random Sample uses uniformly sampled frames as context, while the other baselines retrieve evidence from their long-term memories. For fair comparison, all retrieval-based methods use top-$K = 5$. We evaluate each method with Qwen3.5-35B-A3B (Qwen Team, 2026), Qwen3-VL-30B (Bai et al., 2025), Mistral-Small-3.1-24B (Mistral AI, 2025), and Gemma-3-27B (Gemma Team, 2025) as response models,

Table 3: Comparison of retrieval strategies on EgoLifeQA and Ego-R1 Bench. We report retrieval time per question (Time/Q) and Avg. accuracy on each dataset using Qwen3-VL-30B.

Method	Time/Q (s) ↓	EgoLifeQA (%) ↑	Ego-R1 Bench (%) ↑
Random	0.242	32.4	40.4
BM25	0.253	40.1	47.5
TF-IDF	**0.240**	39.2	47.0
Dense	0.269	**40.5**	**48.5**
Hybrid	0.302	39.3	47.7

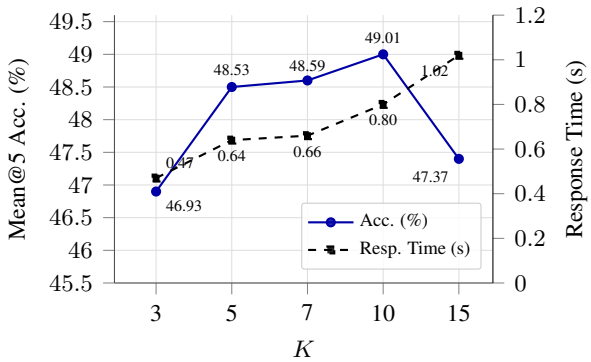


Figure 4: **Effect of retrieval depth K on efficiency and QA performance.** Response time includes retrieval and inference time per question, and accuracy is reported as Mean@5.

repeat each experiment 5 times, and report the average as Mean@5. In task-wise results, Avg. is computed as a sample-size-weighted average across task categories, where each category is weighted by its number of test samples.

Datasets. We evaluate on EgoLifeQA (Yang et al., 2025) and Ego-R1 Bench (Tian et al., 2025), two very-long video understanding benchmarks built on the same days-level video corpus. The corpus records six persons living in a house for seven consecutive days, resulting in continuous and repetitive daily-life video streams. EgoLifeQA contains 500 multiple-choice questions across five tasks, including EntityLog (EL), EventRecall (ER), HabitInsight (HI), RelationMap (RM), and TaskMaster (TM). Ego-R1 Bench contains 2,905 multiple-choice questions and further introduces Future Plan (FP), requiring broader long-term reasoning. Together, these benchmarks evaluate an agent's ability to retrieve sparse evidence and reason over both concrete events and abstract long-term patterns.

More implementation details are provided in Appendix B.

4.1 Main Results

AEMAgent consistently achieves the best overall performance across response models. As shown in Table 1, AEMAgent obtains the highest weighted Avg. accuracy on both EgoLifeQA and Ego-R1 Bench under all four response models. On EgoLifeQA, it outperforms the strongest baseline by 1.9–3.3 percentage points, while on Ego-R1 Bench, the improvement reaches 1.9–9.9 percentage points. These results demonstrate that the proposed active memory formation strategy is effective across different response models.

AEMAgent is particularly effective for long-horizon reasoning tasks. The larger gains on the more challenging Ego-R1 Bench, together with strong performance on HabitInsight, RelationMap, TaskMaster, and Future Plan, show that AEMAgent is well suited to long-horizon reasoning, where models need to retrieve sparse evidence and capture recurring behaviors, relations, routines, and future-oriented patterns. This suggests that active memory formation provides more compact and relevant evidence than fixed-window memory construction.

4.2 Efficiency Analysis

AEMAgent achieves superior memory efficiency. To assess memory cost, Table 2 reports memory tokens and retrieval time per question. AEMAgent achieves the highest accuracy of 48.5% with only 8,777K memory tokens, which is the lowest cost among methods using both episodic and semantic memory. Compared with the second-best WorldMM, AEMAgent uses 32.2% fewer tokens, reduces retrieval time from 6.732s to 0.269s, and improves accuracy by 8.5 percentage points. Although EgoRAG is cheaper, it lacks semantic memory and performs worst with 21.9% accuracy. These results show that AEMAgent builds compact yet informative memories for efficient very-long video reasoning.

4.3 Retrieval Strategy Analysis

Table 3 compares five retrieval strategies with Qwen3-VL-30B as the response model. Random retrieval samples memories without query matching, BM25 (Robertson and Walker, 1994) and TF-IDF (Salton and Buckley, 1988) rely on lexical matching, Dense retrieval ranks memories by Qwen3-Embedding-4B semantic similarity, and Hybrid retrieval combines lexical and dense scores with equal weights.

Table 4: **Ablation study of memory components on Ego-R1 Bench.** Results are reported as Mean@5 accuracy (%). E, S, and V denote episodic, semantic, and visual memory, respectively. "VES w/o QA" uses episodic, semantic, and visual memory but removes the self-generated QA from episodic memory.

Config.	EL	ER	HI	RM	TM	FP	Avg.
E	51.7	43.1	43.2	**49.6**	49.5	44.2	46.4
S	46.8	43.2	43.3	35.5	33.3	35.2	41.9
V	45.8	38.9	41.8	35.5	38.4	33.5	39.8
E+S	52.9	42.8	45.3	48.0	**50.2**	**47.7**	46.6
V+E	54.0	42.2	46.2	45.5	45.8	40.3	45.7
V+S	43.9	39.0	42.0	31.0	35.4	29.0	38.4
VES w/o QA	46.6	38.2	45.1	35.0	49.5	37.9	40.6
VES full	**54.8**	**46.8**	**47.4**	47.3	47.5	43.5	**48.5**

Dense retrieval provides the best accuracy–efficiency trade-off. As shown in Table 3, Dense retrieval achieves the highest Avg. accuracy on both EgoLifeQA and Ego-R1 Bench, reaching 40.5% and 48.5%, respectively. Compared with lexical retrieval, Dense retrieval better captures semantic relevance beyond surface word overlap, which is important for retrieving question-relevant memories in very-long video understanding. Although it is not the fastest strategy, its retrieval time remains low at 0.269s per question, close to TF-IDF and BM25. These results show that semantic retrieval improves QA performance while maintaining practical retrieval efficiency.

4.4 Ablation Study

4.4.1 Effect of Retrieval Depth K_m.

As described in Section 3.6, AEMAgent retrieves top-K_m memories from each textual memory type $m \in \{\text{epi}, \text{sem}\}$. To study the accuracy–efficiency trade-off, we use the same depth for episodic and semantic memories and align retrieved episodic memories with representative frames as visual evidence.

Moderate retrieval depth is sufficient for effective long-horizon reasoning. As shown in Figure 4, increasing K_m from 3 to 10 improves Mean@5 accuracy from 46.93% to 49.01%, indicating that additional memories can provide useful long-term evidence. However, the gain becomes marginal after $K_m = 5$, where AEMAgent already reaches 48.53% accuracy with a response time of 0.64s. Further increasing K_m brings higher latency and even degrades accuracy to 47.37% at $K_m = 15$. This suggests that excessive retrieval may introduce redundant or irrelevant evidence and distract the response model. Therefore, we set $K_m = 5$ as the default retrieval depth to balance accuracy and efficiency.

4.4.2 Effect of Memory Components

AEMAgent maintains three complementary memory sources: visual memory (V) for fine-grained visual grounding, episodic memory (E) for event-level captions and self-generated QA, and semantic memory (S) for consolidated long-term knowledge. To examine their contributions, Table 4 compares single-memory variants, combined-memory variants, the full VES configuration, and a variant without self-generated QA.

Full memory achieves the best and most balanced performance. *VES full* achieves the highest average accuracy of 48.5% on Ego-R1 Bench, showing that visual, episodic, and semantic memories are complementary. While *E+S* benefits long-horizon tasks such as TaskMaster and Future Plan by adding abstract semantic knowledge, it is less effective for event-level questions requiring fine-grained visual details. Incorporating visual memory leads to more balanced performance across task categories.

Self-generated QA is crucial for exploratory episodic memory. Removing QA from the full memory setting causes a large drop from 48.5% to 40.6%. This indicates that simply combining multiple memory sources is insufficient without active event exploration. By asking and answering event-centered questions, AEMAgent enriches episodic memory with more reasoning-oriented evidence, enabling visual, episodic, and semantic memories to work together more effectively.

5 Conclusion

In this paper, we proposed AEMAgent, a cognition-inspired agent that transforms fixed-window passive recording into active memory formation for very-long video understanding. By deciding *when to remember*, *how to explore events*, and *when to reflect*, AEMAgent selectively writes salient events, enriches episodic evidence, updates semantic knowledge from prediction–observation mismatches, and maintains compact long-term memory. Experiments show that AEMAgent delivers consistent gains across response models, especially on long-horizon reasoning tasks, while maintaining a strong efficiency–performance trade-off, highlighting active memory formation as a scalable paradigm for very-long video understanding.

6 Limitations

Although AEMAgent shows strong effectiveness and efficiency, this work still has several limitations that leave room for future exploration. First, our experiments mainly focus on days-level daily-life videos in indoor household scenarios. This setting is suitable for evaluating long-term memory and routine reasoning, but future work can further examine AEMAgent in more diverse environments, such as outdoor egocentric videos and mobile robots. Second, AEMAgent adopts a fixed daily schedule for semantic memory consolidation. While this design is simple and effective, more adaptive consolidation schedules may better fit videos with different event densities. Third, our current study focuses on a single-agent memory setting. Extending AEMAgent to personalized or collaborative memory sharing across multiple embodied agents could be an interesting direction for future work.

References

Richard C Atkinson and Richard M Shiffrin. 1968. Human memory: A proposed system and its control processes. In *Psychology of learning and motivation*, volume 2, pages 89–195. Elsevier.

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*.

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. 2023. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*.

Karl Friston. 2005. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456):815–836.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Xiaohu Huang, Hao Zhou, and Kai Han. 2025. Prunevid: Visual token pruning for efficient video large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19959–19973.

Minsoo Kim, Kyuhong Shim, Jungwook Choi, and Simyung Chang. 2026. Infinipot-v: Memory-constrained kv cache compression for streaming video understanding. *Advances in Neural Information Processing Systems*, 38:138983–139013.

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

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*.

Mistral AI. 2025. Mistral small 3.1. https://mistral.ai/news/mistral-small-3-1. Official release blog.

Jongwoo Park, Kanchana Ranasinghe, Kumara Kahatapitiya, Wonjeong Ryu, Donghyun Kim, and Michael S Ryoo. 2026. Too many frames, not all useful: Efficient strategies for long-form video qa. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3569–3588.

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.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Rajesh PN Rao and Dana H Ballard. 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87.

Aniket Rege, Arka Sadhu, Yuliang Li, Kejie Li, Ramya Korlakai Vinayak, Yuning Chai, Yong Jae Lee, and Hyo Jin Kim. 2026. Agentic very long video understanding. *arXiv preprint arXiv:2601.18157*.

Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 232–241. Springer.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, and 1 others. 2025. Dinov3. *arXiv preprint arXiv:2508.10104*.

Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. 1999. A statistical model-based voice activity detection. *IEEE signal processing letters*, 6(1):1–3.

Baiyang Song, Jun Peng, Yuxin Zhang, Guangyao Chen, Feidiao Yang, and Jianyuan Guo. 2026. Ktv: Keyframes and key tokens selection for efficient training-free video llms. *arXiv preprint arXiv:2602.03615*.

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

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. 2025. Adaptive keyframe sampling for long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29118–29128.

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*.

Endel Tulving and 1 others. 1972. Episodic and semantic memory. *Organization of memory*, 1(381-403):1.

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

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, and 1 others. 2025. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

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.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Zhuo Zhi, Qiangqiang Wu, Wenbo Li, Yinchuan Li, Kun Shao, Kaiwen Zhou, and 1 others. 2025. Videoagent2: Enhancing the llm-based agent system for long-form video understanding by uncertainty-aware cot. *arXiv preprint arXiv:2504.04471*.

A Prompt Templates

To improve reproducibility, we provide the prompt templates used by AEMAgent in the following subsections. These prompts correspond to the main language-model-based modules in our framework, including exploratory episodic memory encoding, prediction-error judging, semantic reflection, semantic memory consolidation, and retrieval-augmented question answering. Each prompt specifies the input context, the expected output fields, and the formatting constraints used in our implementation. During inference, the same templates are applied to all videos and benchmarks without task-specific manual modification.

A.1 Exploratory Episodic Memory Encoding Prompt

Figures 5 and 6 show the prompt templates used for exploratory episodic memory encoding. Given a selected event clip with visual frames, audio, and timestamp metadata, the omni-modal model generates a compact factual caption, self-questioning QA pairs, and a short-term prediction. The caption summarizes visual and audio evidence from a first-person perspective, while the QA pairs encourage active exploration of event participants, scene functions, object usage, and potential event significance.

A.2 Prediction-Error Judge Prompt

Figure 7 shows the prompt template used for prediction-error judging. Given the previous short-term prediction and the current event caption, the model is asked to determine whether the current observation violates the previous expectation. The output contains a binary judgment and a brief explanation of the mismatch. The mismatch explanation is used as evidence for subsequent predictive-coding-guided reflection and semantic memory updating.

Prompts of Caption Extraction

System Prompt

You are a concise multimodal captioning module. Transform the video, audio, and transcript context into a compact factual caption with separate visual, audio, and overall-caption fields. Avoid verbose frame-by-frame visual descriptions and avoid full ASR unless exact wording is essential.

User Prompt

Requirements:
Visual: Briefly summarize the key visible actions, entities, and setting from {EGONAME}'s first-person perspective. Keep this compact and avoid frame-by-frame detail.
Audio: Briefly summarize the key audible cues, including speech content when useful. Do not provide a verbose ASR transcript unless the exact wording is essential.
Overall Caption: Merge visual, audio, and transcript context into one cohesive third-person caption.

Output MUST be strictly in the following format:
[visual](A concise visual summary using '{Person}' as the main subject.)
[audio](A concise audio summary, including important speech meaning or short quotes only when necessary.)
[overall caption](A cohesive caption that combines visual evidence, audio cues, and transcript context.)

Figure 5: Prompt template for caption.

A.3 Semantic Reflection Prompt

Figure 8 shows the prompt template used for semantic reflection. When a prediction error is detected, the model receives the previous factual caption, the previous short-term prediction and the current factual caption. It is then asked to infer high-level semantic updates from the event transition. The output is a set of candidate semantic triples that capture causal relations, state changes, recurring behaviors, or expectation violations supported by the observed evidence. These triples are further checked by a local grounding step before being inserted into the semantic memory graph.

A.4 Semantic Memory Consolidation Prompt

Figure 9 presents the prompt template for semantic memory consolidation. Given the accumulated semantic triples over a long video history, the model is asked to organize them into a compact and consistent long-term semantic graph. Specifically, it merges equivalent relations, removes redundant or conflicting edges, and abstracts repeated event-level observations into stable high-level knowledge such as habits, preferences, routines, and social relations. This consolidation step is performed offline and complements online semantic reflection by reducing memory redundancy and improving the retrievability of long-term knowledge.

B Implementation Details

B.1 AEMAgent implementation

AEMAgent first segments the continuous egocentric video stream using a streaming surprise detector. We sample video frames at a fixed interval and extract visual representations with a frozen DINOv3 encoder. The visual background state is maintained by an exponential moving average with update rate $\alpha = 0.1$. A candidate event boundary is triggered when the visual surprise score exceeds the absolute threshold $\theta_v = 0.1$ and is larger than the relative peak threshold $\rho = 2.5$ over recent surprise history. To avoid over-segmentation, we suppress event clips shorter than $\Delta_{\min} = 3.0$ seconds. In addition, an audio-aware gate estimates speech activity from the trailing audio context using energy-based voice activity cues and rejects boundaries that may interrupt ongoing speech or continuous background sounds. Accepted event clips are saved with their original audio tracks and timestamp metadata. All experiments were conducted on a server equipped with 8 NVIDIA H100 GPUs.

Each selected event is encoded by Qwen3-Omni-30B-A3B through an OpenAI-compatible API. We set the generation temperature to 0.1 and the maximum output length to 4096 tokens for event captioning, self-generated QA, and short-term prediction. The generated captions and QA pairs are stored as episodic memory, while representative frames are stored as visual memory. Semantic memory is represented as a subject-centric directed

Prompts of Generating QA

System Prompt

Step into the shoes of a human user. Instead of just describing the video, think about what a person would find intriguing, confusing, or noteworthy. Generate 3 to 5 high-value Q&A pairs focusing heavily on VIDEO DETAILS.

User Prompt

STRATEGIES FOR HUMAN-LIKE QUESTIONS (Suggestions)
You are encouraged to explore these dimensions to make the questions feel "real":
- Intent & Motivation: Why did the character perform a certain action? (e.g., "Why did he suddenly look at the door?")
- Subtle Interactions: Focus on the relationship between people or objects. (e.g., "Was the person trying to hide the phone when the other entered?")
- Spatial & Environment: Notice things in the periphery. (e.g., "What was that blue object that fell off the table during the commotion?")
- Causality & Prediction: What caused an event or what might happen next? (e.g., "What caused the liquid to spill in the first place?")
- Visual Anomalies: Spot the unexpected. (e.g., "Why is the TV on in the background if no one is watching it?")
CONSTRAINTS
- No Summaries: Do not provide a general description or "Fact" section.
- Evidence-Based: Answers must be derived strictly from the video and audio evidence. If the video doesn't show the answer, state the most logical inference or mention it's "not visible."
- Conversational Tone: Questions should sound like a person investigating, not a technical manual.
You MUST output ONLY pure text in a valid JSON format EXACTLY matching this structure:

```
{
  "curiosity_qna": [
    {"question": "Q1...", "answer": "A1..."},
    {"question": "Q2...", "answer": "A2..."}
  ]
}
```

Figure 6: Prompt template for generating Q&A.

multi-graph implemented with NetworkX, where nodes denote entities, objects, concepts, or behaviors, and edges correspond to typed semantic triples. Prediction-error judging, semantic reflection, and memory consolidation are implemented with Qwen3-VL calls following the prompt templates in Appendix A.

B.2 Retrieval and response generation

For retrieval-augmented question answering, AEMAgent indexes episodic captions, self-generated QA memories, and semantic memory entries. Dense retrieval uses Qwen3-Embedding-4B to encode both queries and memories into normalized embeddings, and ranks candidate memories by dot-product similarity. The retrieval query is constructed from the question, answer choices, and available metadata. Unless otherwise specified, we retrieve top-K_m memories from each textual memory type and use $K_m = 5$ as the default setting. Visual evidence is not retrieved independently; instead, representative frames are temporally aligned with the retrieved episodic memories. The final evidence set, including episodic, semantic, and aligned visual memories, is provided to the response model for answer prediction. We evaluate AEMAgent with Qwen3.5-35B-A3B, Qwen3-VL-30B, Mistral-Small-3.1-24B, and Gemma-3-27B. For QA generation, we use deterministic decoding with temperature 0 and a maximum output length of 8192 tokens.

B.3 Baseline implementation

We compare AEMAgent with Random Sample, EgoRAG, M3-Agent, WorldMM, and EGAgent. Random Sample serves as a non-memory baseline that uniformly samples raw video frames as context. EgoRAG is evaluated with the precomputed EgoLife captions and metadata provided by the benchmark. M3-Agent uses prebuilt graph-based long-term memories and embedding-based retrieval. WorldMM uses episodic, semantic, and visual memories with multi-scale episodic captions, while EGAgent follows an entity-graph-based memory pipeline with transcript and frame-text retrieval. For fair comparison, all methods are evaluated on the same EgoLifeQA and Ego-R1 Bench splits with the same response models. For retrieval-based methods, we use the same top-$K = 5$ evidence setting whenever applicable.

Prompts of Prediction-Error Judge

```
Detect Prediction Error (Surprise) and Extract Key Entities with Query Expansion.
- Previously Predicted Action: {LAST_PREDICTION}
- Actual Current Fact: {CURRENT_FACT}
Instruction:
1. Compare the Actual Current Fact with the Previously Predicted Action. Did the actual event significantly
contradict, deviate from, or fail to fulfill the prediction?
2. Extract up to 5 main entities (e.g., specific people, objects, locations) involved in the 'Actual Current Fact'.
For each entity, also provide 1-2 broader concepts or synonyms (e.g., if extracting 'coffee machine', also
include 'coffee' or 'beverage'). Return a single flattened list containing all original and expanded keywords.

Output MUST be in valid JSON format:
{{
   "is_surprise": true/false,
   "mismatch_reason": "...",
   "extracted_entities": ["entity1", "synonym1", "entity2", "broader_concept2"]
}}
```

Figure 7: Prompt template for prediction-error judging.

B.4 Evaluation protocol

We evaluate on EgoLifeQA and Ego-R1 Bench. EgoLifeQA contains 500 multiple-choice questions, while Ego-R1 Bench contains 2,905 multiple-choice questions and further includes Future Plan. Each experiment is repeated five times, and we report the averaged performance as Mean@5. For task-wise results, the Avg. accuracy is computed as a sample-size-weighted average across task categories, where each category is weighted by its number of test samples.

Prompts of Reflection

Input Data:
- Current Fact (Confirmed): {fact}
- Prior Imagination (Predicted): {imag}
- Mismatch Reason : {reason}
- Previous Fact: {previous_fact}
- Retrieved Semantic Context: {semantic_context}

Instructions:
1. Extract atomic triples: [Subject, Predicate, Object].
2. If a person is doing an action in the Fact, assign "status": "confirmed".
3. If an action is from the Imagination, assign "status": "predicted".
4. Output ONLY a valid JSON list.

Format:
```
[
    {"triple": ["Entity1", "relation", "Entity2"], "status": "confirmed", "time": {time}},
    {"triple": ["Entity1", "will_do", "ActionX"], "status": "predicted", "time": {time}}
]
```

Figure 8: Prompt template for semantic reflection.

Prompts of Semantic Memory Consolidation

System Prompt

You are the cognitive memory consolidation module of an AI agent. It is the end of {CURRENT_DAY}, and the agent is entering "sleep mode" to reflect on the events and knowledge gathered about the person: [{PERSON}].

User Prompt

Here are the recent semantic memories associated with [{SUBJECT}]:
{SEMANTIC_CONTEXT}
Your task is to mimic the human brain during sleep and perform two critical memory operations:
SYNAPTIC PRUNING (Forgetting):
Review the memories and identify any that are now outdated, contradictory, or overly redundant. Pay close attention to the timestamps. If a newer memory explicitly contradicts an older one (e.g., DAY 1: "likes green tea", DAY 2: "hates green tea"), mark the OLDER memory's ID for deletion to resolve the conflict.
MEMORY CONSOLIDATION (Abstraction):
Look for patterns, habits, or repetitive behaviors across the remaining memories and abstract them into generalized, long-term rules. For example, if you see multiple entries like "drank coffee at 9:00" and "made coffee at 14:00", you should abstract a new consolidated memory: predicate="has a habit of drinking", object="coffee".
You MUST output your response strictly in valid JSON format with the following structure:
```
{
"edges_to_delete": [1, 3], // List of integer IDs of the edges to remove. Leave empty [] if nothing needs to be deleted.
"edges_to_add": [
{
"predicate": "the new generalized relation (e.g., likes to drink, works as, frequently uses)",
"object": "the target concept/entity"
}
] // List of new abstracted edges to add. Leave empty [] if no abstraction is needed.
}
```

Figure 9: Prompt template for semantic memory consolidation.