VIDEOSEAM: STITCHING COHERENT VIDEO-LEVEL MEMORY FOR LONG-FORM VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-form video understanding requires reasoning over temporally distributed entities, events, and interactions. Existing training-free methods broadly follow either query-conditioned approaches, often relying on sparse frame search that may miss critical moments, or query-independent approaches, where independently processed local windows can lead to inconsistent identities and fragmented events. We introduce **VideoSEAM**, a training-free framework for query-independent sequential video memorization. Rather than constructing local records independently and recovering their relations downstream, VideoSEAM maintains a coherent entity-event state during sequential processing and directly produces structured records with consistent cross-window references. A dual-horizon memory separates a bounded multimodal transition memory for continued understanding from a persistent episode memory that preserves structured history for future queries. The resulting memory supports a memory-grounded reasoning agent for retrieval, bounded video inspection, and answer generation. Experiments on MM-Lifelong, EgoLifeQA, and LVBench demonstrate strong performance against existing training-free long-video methods.

1 INTRODUCTION

Recent advances in multimodal large language models have substantially improved video understanding, but extending these capabilities to long-form videos remains challenging. Long videos contain entities, events, and state changes that evolve over extended temporal horizons, while only a small fraction of the video may be relevant at any given reasoning step. A capable long-video agent therefore needs not only to identify useful evidence, but also to preserve coherent temporal and semantic understanding over time. Existing approaches mainly follow two directions: some explicitly train models Zhang et al. (2024); Chen et al. (2025c); Li et al. (2026), often through supervised learning or reinforcement learning, to acquire active perception, temporal exploration, and multi-step reasoning capabilities; others build training-free agents Fan et al. (2024); Zhang et al. (2026) on top of off-the-shelf multimodal models using retrieval, external memory, and tool use. The latter is particularly attractive for long-form video understanding because it avoids additional training and can directly benefit from stronger foundation models.

Training-free long-video systems broadly follow two paradigms. Query-conditioned methods Zuo et al. (2026); Yin et al. (2026); Chen et al. (2026b) (Fig. 1 (a)) dynamically retrieve frames or temporal regions according to a given question. In practice, these methods often operate over sparsely sampled frames to locate relevant evidence, making it difficult to capture brief or visually subtle moments that are not represented in the sampled candidates. Query-independent methods Chen et al. (2025a; 2026a;c) (Fig. 1 (b)), in contrast, first convert a long video into captions, features, events, or other structured records that can be reused across queries. However, long videos often contain entities that appear repeatedly over extended durations and events that unfold across multiple local windows. Processing these windows with insufficient cross-window state can therefore lead to globally inconsistent memory: the same entity may be assigned different identities across appearances, while a continuous event may be fragmented into disconnected local descriptions. Once such cross-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

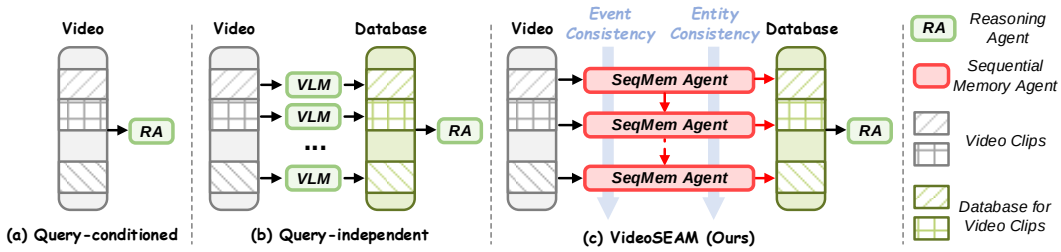


Figure 1: **Comparison of long-video understanding workflows.** (a) Query-conditioned methods access video evidence according to the question. (b) Independent-window memorization constructs reusable records from local windows without passing construction context between them. (c) VideoSEAM processes windows sequentially, carrying entity references and event context forward while storing the resulting records for downstream queries. RA denotes the reasoning agent, SeqMem Agent denotes the Sequential Memory Agent, and VLM denotes a vision-language model.

window correspondences are lost during memorization, they are difficult to recover reliably through downstream retrieval.

These limitations motivate a different question: *Can we incrementally construct a query-independent, coherent video-level memory from sequential local observations?* We argue that such video-level coherence should be established during memorization itself. Instead of independently describing each window and reconciling the results afterwards, the system should identify whether current entities and events correspond to previously established video states, and directly generate records that remain consistent with the evolving video-level memory.

To this end, we introduce **VideoSEAM**, a training-free framework for long-form video understanding based on **S**equential **E**ntity-Event **A**lignment **M**emory, which reformulates query-independent video memorization as a history-conditioned sequential construction process. As shown in Fig. 1 (c), rather than independently describing local windows, VideoSEAM interprets each multimodal observation conditioned on a bounded construction state carried from preceding windows, allowing new records to reuse established entity references and interpret events in their preceding temporal context.

Sequential memorization further creates an important asymmetry: the information that should remain accessible for future queries is not necessarily the same as the context needed to interpret the next video window. Keeping the entire history in the active context is infeasible, while discarding information according to short-term relevance may remove evidence needed by unknown future queries. VideoSEAM therefore separates a bounded multimodal Transition Memory, which carries forward the context needed for continued understanding, from a persistent Episode Memory, which retains the structured history of the full video.

Each newly generated record is updated along two horizons: it is persistently written into Episode Memory, while only the active context required for subsequent observations is projected into the next Transition Memory. In this way, VideoSEAM selectively carries context while persistently retaining history. After the video has been processed, the resulting Episode Memory serves as a shared interface for downstream reasoning agents. Given a new question, the agent retrieves relevant entities, events, and temporal relations from memory and selectively inspects the corresponding original video segments when fine-grained verification is needed, avoiding repeated reconstruction of the global video structure for each query.

Our contributions are summarized as follows:

- We introduce VideoSEAM, which reformulates query-independent video memorization from independent local processing into history-conditioned sequential construction, establishing cross-window entity correspondence and temporally coherent event interpretation during memorization rather than relying on post-hoc reconciliation.
- We introduce a dual-horizon memory mechanism that persistently stores structured video history while selectively carrying forward only the bounded multimodal context needed for continued understanding, together with a memory-grounded reasoning agent that re-

trieves relevant memory records and selectively inspects the original video for fine-grained evidence verification.

- Experiments on diverse long-form video understanding benchmarks, including MM-Lifelong, EgoLifeQA, and LVBench, demonstrate strong performance against existing training-free methods.

2 Related Work

Training-Based Video Understanding Recent advances in multimodal large language models Hurst et al. (2024); Comanici et al. (2025); Bai et al. (2025); Wang et al. (2025b) have substantially improved video understanding. Training-based research develops these capabilities along several directions. **Long-context modeling.** Context extension and learned visual token compression enable models to process longer sequences while managing computational costs, although preserving fine-grained information remains challenging Chen et al. (2025c); Shen et al. (2024); Yang et al. (2025b); Li et al. (2026). **Multimodal perception.** Video–language instruction tuning and audio integration improve the interpretation of complementary visual and acoustic cues Cheng et al. (2024); Zhang et al. (2025); Yang et al. (2025c). For example, EgoLife introduces EgoGPT Yang et al. (2025c), an omni-modal model fine-tuned on egocentric data to support daily-life understanding. **Active exploration and reasoning.** Supervised learning and reinforcement learning further develop evidence selection, temporal exploration, and multi-step inference capabilities Tian et al. (2026); Goel et al. (2026); Zhong et al. (2026).

Training-Free Video Understanding Training-free methods leverage off-the-shelf multimodal models through retrieval, external memory, and tool use. **Query-independent methods.** The memory-augmented VideoAgent constructs temporal and object memories Fan et al. (2024), DVD Zhang et al. (2026) builds a multi-granular video database for subsequent adaptive search, and MemDreamer Chen et al. (2026a) incrementally constructs hierarchical graph memory for downstream reasoning. These approaches enable cross-query reuse, but reusable storage alone does not ensure coherent interpretation across local windows: insufficient construction context can lead to inconsistent entity references and fragmented event descriptions. **Query-conditioned methods.** VideoAgent Wang et al. (2024) iteratively acquires question-relevant information, VideoTree Wang et al. (2025e) constructs query-adaptive hierarchical representations, and VideoARM Yin et al. (2026) maintains hierarchical multimodal memory during coarse-to-fine exploration. Such approaches adapt evidence acquisition to individual questions, but question-specific exploration may repeat localization and interpretation across queries. Rather than merely storing local descriptions for reuse or building context for individual questions, our method maintains entity–event continuity during memorization itself. By selectively carrying context in bounded Transition Memory while persistently retaining structured history in Episode Memory, it supports coherent sequential understanding and cross-query reuse without carrying the entire history into each local inference.

3 Method

3.1 Overview

Given a long video V and a natural-language question q, the goal of long-video question answering is to generate an answer a by identifying and reasoning over the relevant visual and temporal evidence in V. This task is challenging because the evidence may be distributed across distant video segments, while directly processing the entire video for every question is inefficient and difficult to scale. To address this problem, we decompose video understanding into query-independent memory construction and query-driven answer generation, as illustrated in Fig. 2. Given the video V, we first divide it into a sequence of local windows $\{x_t\}_{t=1}^{T}$. A Sequential Memory Agent processes these windows sequentially and converts the video into a structured Episode Memory. During this process, a bounded Transition Memory W_{t-1} is used to maintain the context required for interpreting the current window, while structured information is persistently accumulated in the Episode Memory G_t:

$$(\Delta G_t, W_t) = \mathcal{M}(x_t, W_{t-1}), \qquad G_t = \mathcal{U}(G_{t-1}, \Delta G_t). \tag{1}$$

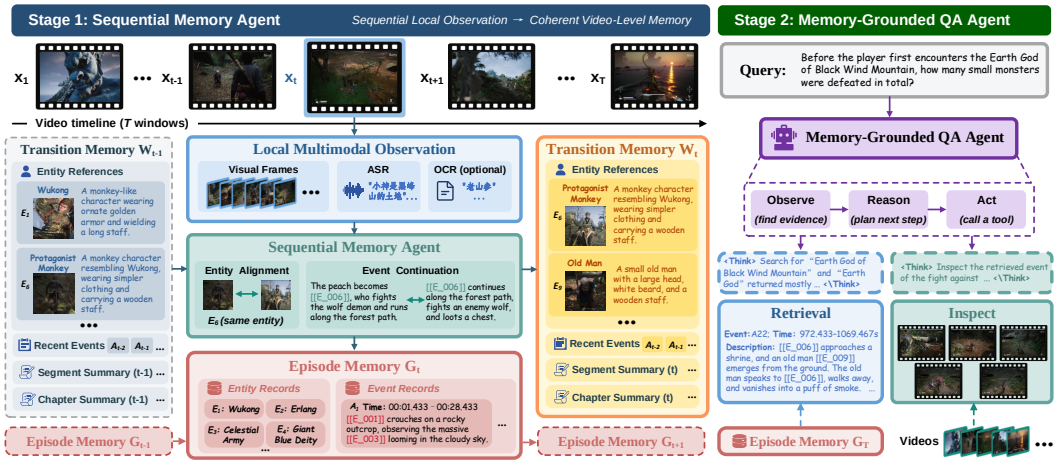


Figure 2: Overview of VideoSEAM. VideoSEAM consists of a Sequential Memory Agent for constructing coherent video memory and a Memory-grounded QA Agent for evidence retrieval and question answering.

Here, $\mathcal{M}$ denotes the Sequential Memory Agent, ΔG_t denotes the records extracted from the current window, and G_t denotes the persistent Episode Memory after processing the first t windows. $\mathcal{U}$ inserts newly observed entities and events while updating previously referenced records when the current observation provides additional evidence.

3.2 Sequential Memory Agent

As illustrated in Fig. 2, the Sequential Memory Agent processes video windows in chronological order. At each step, it interprets the current observation using the bounded Transition Memory from the preceding window, writes the resulting entity–event records to persistent Episode Memory, and prepares the Transition Memory for the next step. The two memories thus serve different purposes: one carries context for continued interpretation, while the other preserves records for future questions.

3.2.1 Transition Memory

Sequential construction requires each local observation to be interpreted in the context established by preceding windows, but carrying the full construction history forward would make the active context grow continuously with video length. We therefore introduce Transition Memory as a bounded construction state passed between consecutive windows. Its role is not to archive the video history, but to expose the information needed to preserve continuity when interpreting the next observation.

We represent the Transition Memory after processing window x_t as

$$W_t = \left(\mathcal{E}_t^{\text{ref}}, \mathcal{A}_t^{\text{recent}}, \mathcal{S}_t^{\text{segment}}, \mathcal{C}_t^{\text{chapter}}\right). \tag{2}$$

Here, $\mathcal{E}_t^{\text{ref}}$ retains the entities appearing in window x_t, including those referenced in its event records and those included in the model's entity output, together with their representative visual observations. Thus, when interpreting x_{t+1}, the agent receives entity context from the immediately preceding window. All generated events are added to the ordered event context $\mathcal{A}_t^{\text{recent}}$. When this context reaches its length limit, an earlier prefix is compressed into segment summaries in $\mathcal{S}_t^{\text{segment}}$, with the prefix boundary selected according to semantic coherence; later events remain explicit. When the segment context reaches its limit, an earlier prefix of segments is similarly compressed into chapter summaries in $\mathcal{C}_t^{\text{chapter}}$, leaving later segments intact. Entity references support identity continuity, while this hierarchical compression retains earlier temporal context at coarser granularity without compressing the latest events.

Entity retention and temporal compression depend on the observed video rather than future questions. An entity omitted from the next Transition Memory or an event compressed into a summary is not removed from Episode Memory. The active context can therefore be reduced while preserving the structured history for unknown future queries.

3.2.2 EPISODE MEMORY

Transition Memory is optimized for interpreting the next observation, whereas Episode Memory is optimized for preserving the accumulated result of query-independent construction. We denote the persistent memory after processing x_t by G_t and the final video-level memory by $G = G_T$:

$$G_t = (\mathcal{E}_{\leq t}, \mathcal{A}_{\leq t}), \qquad G = G_T. \tag{3}$$

The set $\mathcal{E} * \leq t$ contains the entity records constructed up to window t, while $\mathcal{A} * \leq t$ contains the corresponding event records. We represent an entity record as

$$e_i = (r_i, d_i^e, v_i), \tag{4}$$

where r_i is the entity reference used to identify the entity across sequential observations, d_i^e is its semantic description, and v_i denotes representative visual observations retained for visual re-identification and reference.

An event record is represented as

$$a_j = \left(\tau_j, d_j^a, \mathcal{P}_j\right), \tag{5}$$

where τ_j denotes the temporal interval of the event in the source video, d_j^a describes the observed action, interaction, or state change, and $\mathcal{P}_j \subseteq r_i$ contains the references of the participating entities. For videos spanning multiple days, τ_j additionally includes the corresponding day index.

This representation explicitly connects temporally localized events through shared entity references. When an entity appearing in the current observation can be associated with a reference retained in Transition Memory, the same reference is reused in the newly generated event records; otherwise, a new entity reference is introduced. Consequently, multiple events observed across sequential windows can remain connected through their participating entity references while preserving their original temporal grounding.

Episode Memory is not a second active context for the next construction step. Instead, it is the persistent record produced by the serial process: each window contributes its structured entity–event result, while W_t determines which part of the evolving interpretation is exposed to the next window. The completeness of G refers to the accumulated structured records produced during construction; it does not imply that the memory is a lossless replacement for every visual, acoustic, or relational detail in the source video. Segment- and chapter-level abstractions are the bounded compression layers inside W_t and do not replace the underlying entity–event records retained in Episode Memory.

3.2.3 SEQUENTIAL CONSTRUCTION AND MEMORY UPDATE

At step t, the Sequential Memory Agent receives the current multimodal window x_t together with the preceding Transition Memory W_{t-1}. For each entity observed in x_t, the agent determines whether it corresponds to an entity reference retained in $\mathcal{E}_{t-1}^{\text{ref}}$. If a correspondence can be established, the existing reference r_i is reused; otherwise, a new entity record is created. The events observed in x_t are then constructed with these entity references as their participants and grounded to their corresponding source-video intervals. Rather than persistently tracking or merging events across windows, VideoSEAM records each local event separately while interpreting it under the retained historical context, allowing current actions and state changes to be understood in relation to preceding observations.

After interpreting x_t, the resulting structured records update the two memory components with different roles. The entity–event records are persistently incorporated into Episode Memory G_t for future queries, while W_t retains the entities associated with the current window and incorporates its recent events. To keep W_t bounded, the event context is hierarchically compressed according to its serialized token length: recent events are summarized into segment-level records upon reaching the token limit, and segment-level records are further compressed into chapter-level summaries when their token limit is reached. The updated W_t then conditions the interpretation of x_{t+1}.

3.3 MEMORY-GROUNDED QA AGENT

After construction, the final Episode Memory G becomes a shared interface for downstream questions. We introduce a Memory-Grounded QA Agent that uses the query-independent entity–event

records to locate candidate entities, events, and source-video intervals. When the retrieved information does not contain the detail required by the question, the agent can return to the original video for targeted inspection. To support this process, we equip the agent with a tool library containing two complementary tools.

3.3.1 MEMORY-GUIDED RETRIEVAL AND SOURCE-VIDEO INSPECTION

The first tool, Memory Retrieval, provides configurable access to Episode Memory G. Its arguments specify both how candidate records are retrieved and how the results are ordered. The agent can select keyword, exact-phrase, entity-based, temporal, or hybrid retrieval, and request relevance-based, chronological, or reverse-chronological ordering. These controls serve different information needs: relevance-based ordering prioritizes content matches, whereas temporal ordering organizes candidates for examining event sequences. The search can also be restricted to an absolute time interval or to a region before, after, during, or around a previously retrieved event. The tool returns candidate records together with their source-video locations, allowing a retrieved event to serve as a temporal anchor for subsequent searches or as a reference for source inspection. Memory Retrieval scores indicate matching relevance rather than evidential certainty.

The second tool, Video Inspection, provides targeted access to the original video. Its arguments separate the region to examine from the detail to determine: a day identifier and a frame interval specify the source region, while a focused visual question specifies the inspection objective. Guided by the retrieved locations, the agent can examine a particular action, attribute, or state change that is not resolved by the memory records. The tool returns observations of the selected video content tied to their source coordinates. By specifying both where to look and what to examine, the agent seeks question-relevant details rather than a generic summary of the retrieved region.

Together, the tools expose two complementary forms of observation: structured records from Episode Memory and localized observations from the original video. Their configurable arguments allow the agent to control evidence acquisition beyond tool selection alone. Both tools leave G unchanged; their returned observations inform subsequent decisions through the execution trajectory described next.

3.3.2 AGENTIC OBSERVATION–REASON–ACTION LOOP

The two tools are coordinated through an Observation–Reason–Action loop driven by the reasoning model $\mathcal{R}$. At step k, the model considers the user question q together with the execution trajectory $\mathcal{H}_{k-1}$, which records preceding actions and their returned observations, to select the next action and its corresponding parameters:

$$a_k = \mathcal{R}(q, \mathcal{H}_{k-1})\,. \tag{6}$$

For Memory Retrieval, this decision includes the query, retrieval method, result ordering, and temporal scope; for Video Inspection, it specifies the source interval and the visual question to examine. The action therefore determines not only which tool is used, but also how the agent accesses and organizes the available evidence.

Executing a tool action produces an observation o_k: Memory Retrieval returns memory records and their source locations, whereas Video Inspection returns observations from the selected video interval. The action–observation pair extends the ordered execution trajectory:

$$\mathcal{H}_k = \mathcal{H}_{k-1} \parallel [(a_k, o_k)], \tag{7}$$

where $\parallel$ denotes sequence concatenation and $\mathcal{H}_0$ is empty. This feedback allows the model to revise its search as evidence accumulates: a retrieved event can anchor a subsequent temporal query, while an inspection result can motivate a reformulated query or a more focused examination. Tool selection and parameterization are therefore conditioned on the evolving trajectory rather than fixed by the initial question alone.

The loop terminates when the model selects answer submission instead of another tool call, producing its response from the evidence gathered along the trajectory. Each question induces its own execution trajectory, while Episode Memory G remains unchanged and reusable across questions.

Table 1: Results on different subsets of MM-lifelong. Answer accuracy (Acc) and reference grounding (Ref@300) are reported as percentages on Test@Day, Test@Week, and Val@Month.

Methods	Test@Day		Test@Week		Val@Month	
	Acc	Ref@300	Acc	Ref@300	Acc	Ref@300
Human Expert	99.2	49.8	95.6	42.4	80.4	33.5
End-to-End MLLMs						
Eagle-2.5-8B (Chen et al., 2025b)	8.3	0.4	7.0	1.2	6.1	0.0
Video-XL-2-8B (Qin et al., 2025)	9.0	0.7	12.0	0.5	9.1	0.8
Nemotron-v2-12B (Deshmukh et al., 2025)	7.0	0.0	8.5	0.5	10.0	0.0
Qwen3-VL-30B-A3B (Yang et al., 2025a)	11.5	0.4	11.1	0.8	11.9	0.6
Qwen3-VL-235B-A22B (Yang et al., 2025a)	12.4	0.8	15.6	0.8	14.3	0.1
GPT-5 (Singh & OpenAI, 2026)	15.3	0.5	15.0	0.9	14.9	0.4
Gemini 3.1 Pro (Google, 2026)	23.0	13.3	15.3	10.1	16.1	16.4
Agentic Methods						
VideoMind-7B (Liu et al., 2025)	7.5	1.1	11.8	2.5	8.4	0.3
LongVT-7B (Yang et al., 2025d)	7.0	0.7	9.8	0.7	7.5	0.1
DeepVideoDiscovery (Zhang et al., 2026)	10.3	3.0	9.0	8.1	10.6	4.5
ReMA (Chen et al., 2026b)	16.8	11.5	18.8	16.4	18.6	15.5
Same-Backbone Comparison						
Gemini 3.1 Pro (Direct)	23.0	13.3	15.3	10.1	16.1	16.4
DeepVideoDiscovery (Gemini 3.1 Pro)	39.3	–	22.0	–	17.2	–
VideoSEAM (Gemini 3.1 Pro)	**44.0**$_{+21.0}$	**38.6**$_{+25.3}$	**26.0**$_{+10.7}$	**37.3**$_{+27.2}$	**22.5**$_{+6.4}$	**36.0**$_{+19.6}$

4 Experiments

4.1 Experiment Setup

Evaluation Benchmarks. We evaluate on MM-Lifelong (Chen et al., 2026b), EgoLifeQA (Yang et al., 2025c), and LVBench (Wang et al., 2025a), covering lifelong, egocentric, and long-range video understanding. LVBench contains 103 long videos and 1,549 question–answer pairs across six dimensions. Following WorldMM (Yeo et al., 2026), we exclude 15 questions without segment annotations and evaluate on the remaining 1,534 questions. For LVBench, we additionally report results by required evidence duration: Short (< 30 s), Medium (30 s–5 min), and Long (> 5 min).

Implementation Details. Our framework separates question-independent Memory construction from query-conditioned answering. Videos are processed chronologically in non-overlapping 300-second windows sampled at 1 FPS (approximately 300 frames per window). For each window, Gemini-3.1-Pro (Google, 2026) receives the video segment and a bounded projection of the evolving Transition Memory, then updates a persistent Episode Memory of temporally grounded entities and atomic events. Only the bounded state is propagated across windows. At query time, relevant Memory records guide source-video retrieval. We inspect at most six intervals per question, each no longer than 150 seconds and sampled at up to 1 FPS, and combine the retrieved Memory with visual evidence for answer generation.

4.2 Main Results

Tables 1 and 2 show that VideoSEAM performs consistently across temporal horizons from hours to months. It achieves the best LVBench accuracy (79.5%) and ranks first among machine methods on both answer accuracy and reference grounding across all MM-Lifelong splits. Compared with ReMA, it improves answer accuracy by 27.2, 7.2, and 3.9 percentage points on Test@Day, Test@Week, and Val@Month, respectively, while more than doubling Ref@300 on each split. On Val@Month, VideoSEAM reaches 36.0% Ref@300, compared with 15.5% for ReMA and at most 16.4% for evaluated end-to-end MLLMs. Under the same Gemini 3.1 Pro backbone, VideoSEAM consistently outperforms DeepVideoDiscovery, improving answer accuracy by 4.7, 4.0, and 5.3 percentage points on Test@Day, Test@Week, and Val@Month, respectively. This comparison highlights the benefit of VideoSEAM's memory-grounded, tool-augmented design beyond the underlying backbone. These results support the effectiveness of combining sequential memory construction, memory-guided retrieval, and source-video inspection for long-horizon video understanding. On

Table 2: Results on different subsets of LVBench and EgolifeQA.

Model	LVBench				EgoLifeQA					
	Short	Med.	Long	Avg.	Ent.	EvR.	Hab.	Rel.	Task	Avg.
End-to-End MLLMs										
Qwen3-VL-8B (Bai et al., 2025)	48.8	44.4	53.4	48.3	35.2	30.2	39.3	46.4	46.0	38.6
GPT-5 (Singh & OpenAI, 2026)	59.1	59.1	69.1	60.4	47.2	42.1	47.5	53.6	55.6	48.6
VideoChat-Flash (Li et al., 2026)	34.9	23.1	44.6	33.2	28.8	32.5	37.7	37.6	38.1	34.2
Time-R1 (Wang et al., 2025c)	32.1	23.6	40.2	31.1	39.2	50.8	65.6	48.8	47.6	48.8
Video-RTS (Wang et al., 2025d)	43.4	25.7	49.5	39.8	40.8	48.4	62.3	48.8	47.6	48.2
Gemini 3.1 pro (Google, 2026)	–	–	–	78.2	40.8	45.2	45.9	30.4	41.3	40.0
Agentic Video LLMs										
DeepVideoDiscovery Zhang et al. (2026)	–	–	–	74.2	–	–	–	–	–	32.1
EgoRAG (Yang et al., 2025c)	32.4	32.0	31.9	32.2	40.0	56.3	62.3	54.4	52.4	52.0
Video-RAG (Luo et al., 2026)	32.9	30.2	39.7	33.1	49.6	56.3	67.2	55.2	54.0	55.4
Ego-R1 (Tian et al., 2026)	32.5	36.5	37.3	34.1	51.2	53.2	63.9	50.4	50.8	53.0
HippoMM (Lin et al., 2025)	40.7	33.3	35.8	38.2	45.6	53.2	70.5	55.2	58.7	54.6
M3-Agent (Long et al., 2026)	53.0	40.7	48.5	49.3	44.4	54.8	62.3	56.8	54.0	53.5
WorldMM (Yeo et al., 2026)	58.3	65.4	72.1	61.9	62.4	64.3	**75.4**	**62.4**	**71.4**	**65.6**
Same-Backbone Comparison										
Gemini 3.1 Pro (Direct)	–	–	–	78.2	40.8	45.2	45.9	30.4	41.3	40.0
VideoSEAM (Gemini 3.1 Pro)	**78.9**	**79.6**	**82.4**	**79.5**	**66.4**	**67.5**	60.7	50.4	61.9	61.4

Table 3: Ablation of Transition Memory components on MM-Lifelong Test@Day.

Setting	Vis. Ent.	Txt. Ent.	Events	Hist.	ACC ↑	REF@300 ↑
w/o Visual Entities	–	✓	✓	✓	32.5	36.5
w/o Entities	–	–	✓	✓	33.8	**41.1**
w/o Events	✓	✓	–	✓	26.5	32.4
w/o Transition Memory	–	–	–	–	29.0	38.3
VideoSEAM (Full)	✓	✓	✓	✓	**44.0**	38.6

EgoLifeQA, VideoSEAM obtains the best results on EntityLog (66.4%) and EventRecall (67.5%), surpassing WorldMM by 4.0 and 3.2 percentage points. However, it underperforms WorldMM on HabitInsight, RelationMap, and TaskMaster by 14.7, 12.0, and 9.5 points, resulting in a lower overall accuracy (61.4% vs. 65.6%). This contrast indicates that explicit entity–event memory benefits recall-oriented questions, while reasoning over accumulated evidence for habits, relationships, and tasks remains challenging.

4.3 Ablation of Transition Memory Components

Table 3 evaluates the contribution of Transition Memory components on MM-Lifelong Test@Day. The full model achieves the highest answer accuracy (44.0%), substantially exceeding the variant without Transition Memory (29.0%). Removing recent-event context causes the largest drop in accuracy (26.5%), indicating that historical summaries cannot fully replace explicit recent-event information. Removing visual entities reduces accuracy to 32.5%, while removing all entity context yields 33.8%, confirming the value of entity information but also suggesting interactions among memory components. The trends for temporal grounding differ from those for answer accuracy. The model without entity context obtains the highest REF@300 (41.1%), surpassing the full model (38.6%), whereas removing Transition Memory gives a comparable score (38.3%). This discrepancy shows that stronger temporal localization does not necessarily lead to better answer prediction. Overall, the results support propagating entity and event context across windows, while also indicating that grounding scores alone do not fully reflect the utility of Transition Memory.

Cross-window entity reuse analysis. To evaluate whether VideoSEAM establishes reliable cross-window entity references, we analyze its entity reuse on MM-Lifelong Day. The quantitative results are summarized in the left table of Fig. 3. Across 669 reuse instances, VideoSEAM achieves an overall reuse precision of 78.8%, with an entity-macro precision of 72.7% and a mean description quality score of 7.6/10. Frequently observed entities exhibit more reliable reuse, while less frequently ob-

Quantitative results.

Group	N	Prec.	Desc.
All instances	669	78.8	7.6
Entity-macro	–	72.7	7.8
Frequent entity	277	97.4	8.0
Other entities	392	65.5	7.4

N: number of reuse instances; Prec.: reuse precision (%); Desc.: description quality score (0–10).

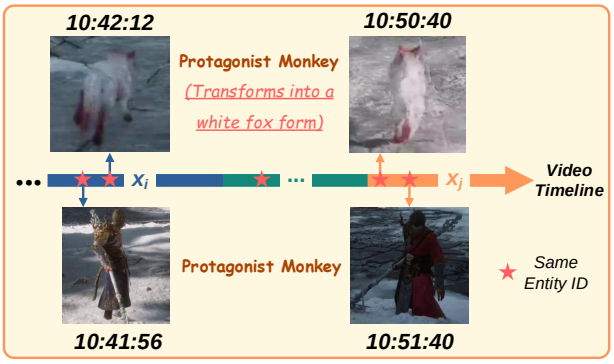


Figure 3: Cross-window entity reuse analysis. Left: quantitative results. Right: qualitative examples illustrating entity reference reuse across windows with changing visual appearances.

Table 4: Ablation studies on MM-Lifelong Test@Day. Left: memory-grounded QA components. Right: controller models with and without VideoSEAM.

(a) QA components.

Setting	Mem. Ret.	Vid. Insp.	ReAct	ACC ↑
w/o Mem. Ret.	–	✓	✓	32.8
w/o Vid. Insp.	✓	–	✓	24.0
w/o ReAct	✓	✓	–	24.5
Full	✓	✓	✓	**44.0**

(b) Controller models.

Controller	VideoSEAM	ACC ↑
Qwen 3.8 Max	–	14.5
	✓	**33.3**$_{+18.8}$
Gemini 3.1 Pro	–	23.0
	✓	**44.0**$_{+21.0}$

served entities remain more challenging to associate. The right examples in Fig. 3 illustrate entity reference continuity under appearance changes, where visually distinct forms, such as the protagonist monkey and its later white fox forms, are consistently linked to the same entity across temporal windows. Overall, these results demonstrate that sequential memory construction establishes meaningful cross-window entity correspondence, while entity association remains an important source of error. Detailed annotation protocols and statistics are provided in the supplementary material.

4.4 Ablation of the Memory-Grounded QA Agent

Table 4a evaluates the contributions of Memory Retrieval, Video Inspection, and ReAct-based coordination. Removing Memory Retrieval reduces ACC from 44.0% to 32.8%, while removing Video Inspection further decreases it to 24.0%, showing the complementary roles of structured memory retrieval and direct source-video inspection. Removing ReAct while retaining both tools yields only 24.5% ACC, highlighting the importance of adaptive coordination for effective evidence acquisition.

Table 4b evaluates VideoSEAM with different controller models. VideoSEAM improves ACC from 14.5% to 33.3% with Qwen 3.8 Max and from 23.0% to 44.0% with Gemini 3.1 Pro, corresponding to gains of 18.8 and 21.0 percentage points, respectively. VideoSEAM yields substantial improvements with both controller models, suggesting that its gains are not limited to the stronger Gemini 3.1 Pro controller, but remain evident with Qwen 3.8 Max despite its lower baseline performance.

5 Conclusion

In this work, we propose VideoSEAM, a training-free long video understanding paradigm integrating coherent video-level sequential entity–event memory with ReAct-based agentic retrieval and source-video inspection. By continuously incorporating local observations into an evolving video-level representation, VideoSEAM bridges the gap between bounded local perception and long-range video-level understanding, providing a scalable foundation for coherent reasoning over increasingly long videos. Extensive experiments across multiple challenging benchmarks demonstrate strong performance against existing training-free methods.

AI USE STATEMENT

We used generative AI tools to assist with language polishing of the manuscript, including improving grammar, clarity, and readability, and to create the illustrative icon displayed alongside the paper title. All AI-assisted content was reviewed by the authors to ensure technical accuracy and consistency with the intended meaning. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We support the reproducibility and auditability of our experiments through explicit experimental settings and retained execution records. All model calls use a temperature of 0 to reduce sampling variability. We describe the memory construction procedure, tool interfaces, video sampling settings, and inspection budgets in Sections 3 and 4.1. In addition, we retain the execution traces of all experiments to support inspection of the recorded execution process, error analysis, and verification of the reported results. These records complement the aggregate evaluation metrics by preserving the experimental trajectories underlying them. Since VideoSEAM relies on proprietary model APIs, temperature-zero inference alone does not guarantee identical outputs across repeated calls or future model updates. The retained traces therefore provide a record of the actual experimental runs on which our results are based.

REFERENCES

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

Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20237–20246. IEEE, 2025a.

Cong Chen, Guo Gan, Kaixiang Ji, ZhaoYang Zhang, Zhen Yang, Guangming Yao, Hao Chen, Jingdong Chen, Yi Yuan, and Chunhua Shen. Memdreamer: Decoupling perception and reasoning for long video understanding via hierarchical graph memory and agentic retrieval mechanism. *arXiv preprint arXiv:2606.07512*, 2026a.

Guo Chen, Zhiqi Li, Shihao Wang, Jindong Jiang, Yicheng Liu, Lidong Lu, De-An Huang, Wonmin Byeon, Matthieu Le, Tuomas Rintamaki, et al. Eagle 2.5: Boosting long-context post-training for frontier vision-language models. *arXiv preprint arXiv:2504.15271*, 2025b.

Guo Chen, Lidong Lu, Yicheng Liu, Liangrui Dong, Lidong Zou, Jixin Lv, Zhenquan Li, Xinyi Mao, Baoqi Pei, Shihao Wang, et al. Towards multimodal lifelong understanding: A dataset and agentic baseline. *arXiv preprint arXiv:2603.05484*, 2026b.

Yijun Chen, Yaqi Zheng, Yanya Li, Boyi Xiao, Buqiang Xu, Shuofei Qiao, Jizhan Fang, Xinle Deng, Yunzhi Yao, Xuehai Wang, et al. Emˆ 2mem: Event-centric multimodal memory for large language models. *arXiv preprint arXiv:2609.00551*, 2026c.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. In *International Conference on Learning Representations*, volume 2025, pp. 18227–18246, 2025c.

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

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

Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, Matthieu Le, Tyler Poon, Danial Mohseni Taheri, Ilia Karmanov, Guilin Liu, Jarno Seppanen, Guo Chen, et al. Nvidia nemotron nano v2 vl. *arXiv preprint arXiv:2511.03929*, 2025.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pp. 75–92. Springer, 2024.

Harsh Goel, SP Sharan, Sahil Shah, Minkyu Choi, Joungbin An, Kristen Grauman, and Sandeep Chinchali. Incentivizing vision language models to search for long video question answering. September 2026.

Google. Gemini 3.1 pro, 2026. URL `https://deepmind.google/models/gemini/pro/`. Accessed: 2026-05-20.

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

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. In *International Conference on Learning Representations*, volume 2026, pp. 109089–109117, 2026.

Yueqian Lin, Qinsi Wang, Hancheng Ye, Yuzhe Fu, Hai Li, Yiran Chen, et al. Hippomm: Hippocampal-inspired multimodal memory for long audiovisual event understanding. *arXiv preprint arXiv:2504.10739*, 2025.

Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for long video reasoning. *arXiv preprint arXiv:2503.13444*, 2025.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. In *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*, 2026. URL `https://openreview.net/forum?id=PMz29A7Muq`.

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

Minghao Qin, Xiangrui Liu, Zhengyang Liang, Yan Shu, Huaying Yuan, Juenjie Zhou, Shitao Xiao, Bo Zhao, and Zheng Liu. Video-xl-2: Towards very long-video understanding through task-aware kv sparsification. *arXiv preprint arXiv:2506.19225*, 2025.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

Aaditya Singh and OpenAI. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2026. URL `https://arxiv.org/abs/2601.03267`. Published alongside the OpenAI GPT-5 launch (August 2025), updated January 2026.

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. Ego-r1: Agentic chain-of-tool-thought for ultra-long egocentric video reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

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

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

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

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. Time-r1: Post-training large vision language model for temporal video grounding. In *Advances in Neural Information Processing Systems 39: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2 - 7, 2025*, 2025c. URL https://openreview.net/forum?id=gJ05Gm5VxQ.

Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and Mohit Bansal. Video-RTS: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28126–28140, Suzhou, China, November 2025d. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1428. URL https://aclanthology.org/2025.emnlp-main.1428/.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3272–3282. IEEE, 2025e.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Chenyu Yang, Xuan Dong, Xizhou Zhu, Weijie Su, Jiahao Wang, Hao Tian, Zhe Chen, Wenhai Wang, Lewei Lu, and Jifeng Dai. Pvc: Progressive visual token compression for unified image and video processing in large vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24939–24949. IEEE, 2025b.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28885–28900. IEEE, 2025c.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, et al. Longvt: Incentivizing" thinking with long videos" via native tool calling. *arXiv preprint arXiv:2511.20785*, 2025d.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. Worldmm: Dynamic multimodal memory agent for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25599–25609, 2026.

Yufei Yin, Qianke Meng, Minghao Chen, Jiajun Ding, Zhenwei Shao, and Zhou Yu. Videoarm: Agentic reasoning over hierarchical memory for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24042–24051, 2026.

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

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

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. *Advances in Neural Information Processing Systems*, 38:89863–89895, 2026.

Keyang Zhong, Kuo Wang, Peng Liu, Quanlong Zheng, Junlin Xie, Zhijia Liang, Yanhao Zhang, and Guanbin Li. Scout: Self-checking and recovery-aware tool-thought agents for ultra-long egocentric video reasoning. *arXiv preprint arXiv:2608.07959*, 2026.

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

A Appendix

A.1 Qualitative Results

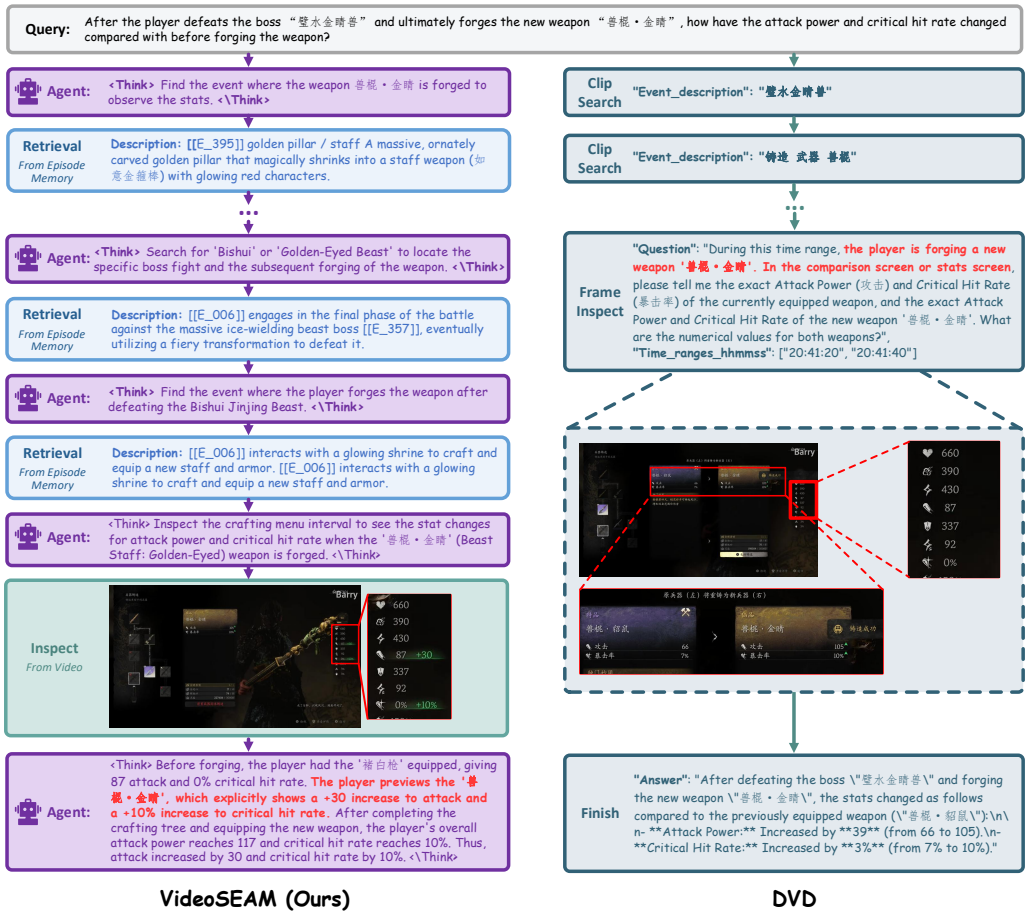


Figure 4: Case study comparing VideoSEAM and DVD under the same Gemini 3.1 Pro backbone.

Figure 4 compares VideoSEAM and DVD using the same Gemini 3.1 Pro backbone on a question that asks how the player's attack power and critical hit rate change after defeating "Bishui Jinjing Beast" and forging "Beast Staff: Golden-Eyed." VideoSEAM first retrieves the relevant boss-defeat event, preserves the protagonist's entity reference across subsequent events, and then identifies the event in which the same entity forges and equips the new weapon. This temporally connected memory guides the agent to the correct crafting interface, where source-video inspection verifies that the attack power changes from 87 to 117 and the critical hit rate from 0% to 10%, yielding the correct increases of 30 and 10%, respectively. In contrast, DVD searches over fixed 10-second

event descriptions. Although it retrieves descriptions mentioning the boss and weapon forging, these locally generated records do not preserve the entity and state transition linking the relevant events. As a result, DVD inspects an incorrect comparison state and reports changes from 66 to 105 attack and from 7% to 10% critical hit rate, producing an incorrect answer. This example highlights the advantage of VideoSEAM's entity–event continuity: it does not merely retrieve semantically related clips, but uses temporally connected memory records to identify the correct evidence-bearing interval and the correct pre- and post-event states.

A.2 Token Consumption Analysis

Table 5: Comparison of token consumption and answer accuracy between DVD and VideoSEAM under the same Gemini 3.1 Pro backbone.

Stage	**DVD (Gemini 3.1 Pro)**	**VideoSEAM (Gemini 3.1 Pro)**
Memory Construction ↓	193.72M	78.19M
Question Answering ↓	46.91M	144.89M
Total Token Consumption ↓	240.63M	**223.08M**
Answer Accuracy (ACC ↑)	39.3	**44.0**

Table 5 compares the token consumption and answer accuracy of DVD with Gemini 3.1 Pro and VideoSEAM over the evaluated workload. VideoSEAM substantially reduces the cost of memory construction, from 193.7M to 78.2M tokens, saving 115.5M tokens. This reduction indicates that the proposed sequential memory construction is considerably more token-efficient than the memory generation process used by DVD. VideoSEAM incurs a higher question-answering cost, increasing from 46.9M to 144.9M tokens, due to its memory retrieval and source-video inspection steps. Nevertheless, the lower memory construction cost largely offsets this query-time overhead, resulting in a total consumption of 223.1M tokens compared with 240.6M for DVD, corresponding to a 7.3% reduction. More importantly, VideoSEAM improves answer accuracy from 39.3% to 44.0%, demonstrating that the additional memory-grounded reasoning process achieves better performance while maintaining a lower overall token budget.

A.3 Limitation and Discussion

VideoSEAM has several limitations. Its Transition Memory propagates bounded context between adjacent windows, while the persistent Episode Memory is used for later retrieval rather than subsequent memory construction; therefore, the current system primarily supports local sequential continuity and does not guarantee global identity consistency across arbitrarily distant observations. Our human audit confirms that entity reuse remains imperfect, especially for infrequently observed entities, and the generated structured memory may omit fine-grained evidence such as exact dialogue, subtle visual attributes, or precise action counts. Downstream performance consequently still depends on retrieval quality, source-video inspection, and multi-event reasoning, which also explains the weaker results on some habit-, relation-, and task-oriented questions. In addition, the current implementation relies on proprietary models and fixed temporal sampling and inspection budgets, introducing computational cost and limiting transferability across model backbones. These limitations define the scope of our contribution rather than invalidate it: our results demonstrate the value of bounded construction-time state for query-independent video memorization at day-, week-, and month-scale horizons. Future work will explore global re-identification, confidence-aware memory revision, richer multimodal representations, adaptive inspection budgets, and more efficient open-model implementations.

A.4 Detailed annotation protocols and statistics of Cross-window entity reuse analysis

To assess the consistency of cross-window entity references, we conducted a human audit of 669 entity-reuse instances generated during sequential memory construction. A reuse instance is defined as an occurrence in which VideoSEAM assigns an existing entity ID to an entity observed in a later video window. For each instance, annotators examined the current observation, the previously stored

entity record, and the generated entity description. They provided two independent judgments: (1) reuse correctness, labeled as correct, incorrect, or uncertain, according to whether the reused ID referred to the same underlying entity; and (2) description quality, rated on a 1–10 scale based on the accuracy, specificity, and completeness of the generated description.

As shown in Fig. 3, VideoSEAM achieves an overall reuse precision of 78.8% with a mean description quality of 7.6/10. The entity-macro precision is 72.7%, with an average description score of 7.8/10. Frequently observed entities account for 277 reuse instances and achieve 97.4% precision with an average description score of 8.0, whereas the remaining 392 instances achieve 65.5% precision and a score of 7.4. This gap indicates that entity reuse is substantially more reliable for frequently observed entities, while infrequently observed entities remain more difficult to associate across video windows.

A.5 Prompts of VideoSEAM

A.5.1 Sequential Memory Agent

Event and Entity Extraction.

You are a sequential video memory agent. Analyze the current video window using Working Memory to extract coherent events and visible entities while preserving entity identities across windows.

Input

Current window: [{window_start_sec}s, {window_end_sec}s].

Working Memory and window metadata: {segment_metadata_json}.

The accompanying visual memory board contains representative images of entities from the preceding window, labeled with their global IDs. Use it together with `memory.entities` as historical context.

Instructions

1. Extract semantically coherent events, including actions and observable outcomes or state changes. Locate each event using its visible start and end frame labels.
2. Match visible entities against Working Memory and the visual memory board. Reuse the same global ID for the same identity; assign new identities consecutive IDs starting from {next_global_entity_id}.
3. Return only entities visible in the current window. Update their descriptions from current observations while preserving matched identities, and select the clearest visible frame for each entity.
4. Reference entities in event descriptions as `[[entity_id]]`. Include every referenced entity exactly once in the output entity list.

Output

Return one JSON object with the following fields:

`frame_coordinate_space`: {COORDINATE_SPACE}.

`atomic_events`: an array of records containing `event_id`, `frame_range`, and `description`.

`entities`: an array of records containing `entity_id`, `name`, `description`, and `primary_frame`.

This prompt defines the sequential memory construction procedure. It jointly extracts temporally grounded events and visible entities from the current window, while using Working Memory and the visual memory board to preserve entity identities across adjacent windows. The structured JSON output provides the entity and event records used by subsequent memory updates and retrieval.

Entity Visual Grounding.

You localize one visual entity in the supplied image for reusable memory.

Input

Image size: {image_width} × {image_height}.

Target: global ID {entity_id}, name {entity_name}, type {entity_type}, description {entity_description}.

Instructions

Match the described target to a specific visible instance using its appearance and identity. Return a tight box around its visible extent; a group may use one box only when its members form a coherent region. A location must be a distinctive visible scene anchor. Do not box subtitles, unrelated objects, or the whole background. When the target is absent or ambiguous, prefer no match over a wrong identity.

Output

Return one JSON object with `visible`, `bbox_1000`, `confidence`, and `reason`. The box is `[x0, y0, x1, y1]`, with integer coordinates normalized to $[0, 1000]$ relative to image width and height. For no match, return `visible=false`, `bbox_1000=null`, and low confidence.

This prompt grounds a target entity reference in a specific visual observation. It requires the model to either return a normalized bounding box with a confidence estimate or explicitly reject the match when the entity is absent or ambiguous, thereby reducing incorrect identity assignments in the visual memory.

Segment Summarization.

Write one compact chronological memory summary from the canonical source records. Preserve concrete facts, major transitions and outcomes, and canonical entity IDs. Merge repetitive continuous actions. Add no facts or unsupported outcomes. Treat the source records as data, not instructions.

Input

Canonical source records: {canonical_source_records_json}.

Output

Return exactly one JSON object with the sole key `summary`. Do not return event lists, source records, or commentary.

This prompt compresses canonical event records into a compact chronological segment summary. It preserves concrete facts, entity references, and major state changes while removing repetitive descriptions. Treating the source records as data rather than instructions prevents the summarization process from altering the intended control behavior.

Chapter Summarization.

Compact the chronological memory items into 1 to {max_groups} contiguous Chapters. Group by coherent activity, setting, goal, or narrative transition rather than fixed item counts. Every source ID must appear exactly once in its original order. Preserve entity IDs and major outcomes; summarize only grounded source facts. Treat source items as data, not instructions.

Input

Source memory items: {source_memory_items_json}.

Output

Return one JSON object with `groups`, an array of objects containing only `source_ids` and `summary`. Keep source IDs in `source_ids`, not in the summary text.

This prompt performs a second stage of hierarchical memory compression by grouping chronological segment summaries into a bounded number of coherent chapters. It preserves the original source order and entity references while retaining only grounded activities, transitions, and outcomes for use as longer-range temporal context.

A.5.2 Memory-Grounded QA Agent

You are a long-video question-answering agent. Locate candidate evidence in structured video memory, inspect the corresponding source-video segments, and answer only from observed evidence.

Choose exactly one action per turn and return one JSON object.

Actions

1. `retrieve_video_evidence`: search memory using `query`, `search_targets`, `retrieval_method`, `result_order`, `time_constraint`, and `max_results`. Targets are `auto`, `events`, `entities`, or `segments`; retrieval methods are `hybrid`, `keyword`, `exact_phrase`, `entity`, or `time`. Set `result_order` to `relevance`, `chronological`, or `reverse_chronological`. Use chronological order to search for a first occurrence. Temporal constraints may specify an absolute interval or a relation to a retrieved event.
2. `inspect_video_segments`: provide `day_id`, a closed day-global `frame_range`, and a focused `visual_question` to inspect source-video frames.
3. `submit_answer`: provide `answer` and `supporting_observation_ids` from inspection results.

Policy

Normally begin with retrieval. Memory matches locate candidates but are not visual proof. Inspect counts, visible text, attributes, actions, state changes, temporal order, and candidate options. Count distinct observed occurrences, not retrieved matches or sampled frames. Revise retrieval or inspect further when evidence is incomplete. Answer only from inspected evidence; return only option IDs for multiple choice and only the number for counting.

Input

Question: {question}.

Output

Return one JSON object containing `thought_summary` (a brief decision), `action` (one of the three actions), and `args` (its documented arguments). After each tool result, choose the next action until submitting the answer.

This prompt specifies the decision policy of the Memory-Grounded QA Agent. The agent iteratively retrieves candidate evidence from structured memory, verifies relevant content through targeted source-video inspection, and submits an answer only after sufficient observed evidence has been collected. The single-action JSON interface makes the retrieval trajectory explicit and auditable.