VIDEOTAPESTRY: QUERY-ADAPTIVE MEMORY REFINEMENT FOR MULTI-AGENT LONG-VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-video understanding places substantial demands on memory, as answering questions often requires retrieving information distributed across extended temporal spans. Existing approaches broadly follow two paradigms: query-driven exploration, which is sensitive to localization errors, and query-independent memory construction, which may omit question-specific details. We introduce **VideoTapestry**, a training-free multi-agent framework that adapts a preconstructed hierarchical video memory through coarse-to-fine, query-driven refinement. The preconstructed memory organizes video content into three levels, capturing global narrative context, event-level temporal structure, and fine-grained relational evidence, respectively. To support coarse-to-fine localization and observation, we assign a specialized agent to each level, keeping retrieval and refinement within a scale-specific context. Guided by the query, these agents revisit relevant video regions and enrich layer-wise memories with targeted multimodal observations. Their refinements are assembled according to the original hierarchy into a composite query-adaptive memory, preserving global context in a compact form while retaining fine-grained evidence along query-relevant branches for final reasoning. Compared with direct GPT-5.5 inference, VideoTapestry achieves absolute accuracy gains of 17.2%, 14.9%, 9.8%, and 7.0% on LVBench, LongVideoBench (Long), Video-MME (Long), and EgoSchema, respectively, achieving the state-of-the-art results among all competitors.

1 INTRODUCTION

Long-video understanding is a fundamental capability for vision-language models (VLMs) reasoning over extended activities, narratives, and real-world interactions (Fu et al., 2025; Wu et al., 2024; Wang et al., 2025a; Mangalam et al., 2023). Compared with understanding short clips, answering questions about hour-long videos requires a model to preserve global context while locating decisive evidence at the appropriate semantic granularity. Uniform sampling may miss brief events, while dense frame sampling introduces substantial redundancy and query-irrelevant information, making relevant content difficult to isolate (Ye et al., 2025; Wang et al., 2025b; Wu et al., 2024). These challenges motivate memory mechanisms that organize video information for selective evidence retrieval while preserving the broader context needed for reasoning.

Existing approaches can be broadly categorized into two paradigms. As shown in Figure 1(a), top-down methods start from the query and progressively inspect relevant intervals, constructing or updating hierarchical memory during the search (Wang et al., 2025b; Ye et al., 2025; Yin et al., 2026; Lin et al., 2026a). This enables adaptive observation, but inaccurate localization may direct the search away from the key video segments. Bottom-up methods, as illustrated in Figure 1(b), construct a query-independent video memory and retrieve or revisit relevant content at inference time (Fan et al., 2024; Zuo et al., 2026; Zhang et al., 2026; Yeo et al., 2026). Such representations provide broad coverage but may omit details required by specific questions. Intuitively, these two paradigms are complementary: query-independent memory provides broad coverage, while query-driven exploration acquires question-specific evidence.

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.

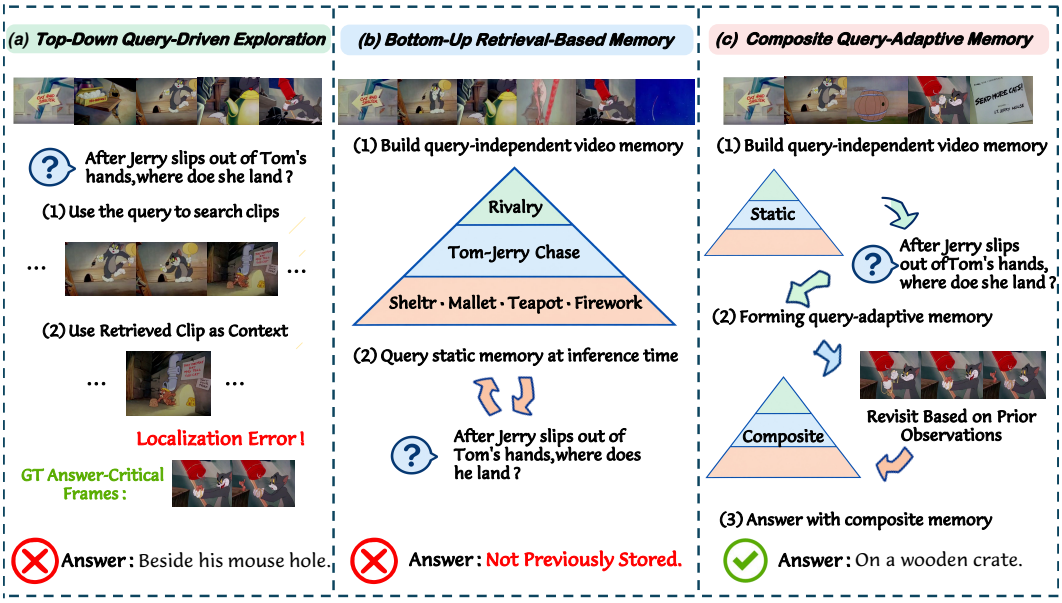


Figure 1: **Comparison of long-video context construction paradigms. Left:** Top-down exploration may miss answer-critical evidence due to localization errors. **Middle:** Bottom-up memory provides broad coverage but may omit query-specific details. **Right:** VideoTapestry combines a query-independent hierarchy with targeted revisiting to form a composite query-adaptive memory that preserves global context and enriches relevant evidence.

However, integrating query-specific evidence with existing memory while preserving global context remains challenging. Integrating the strengths of the two paradigms requires incorporating new evidence into memory while limiting irrelevant content. This process involves global localization, temporal narrowing, and fine-grained verification. These complementary reasoning objectives motivate specialized agents that refine memory at corresponding semantic levels and pass inspection objectives between levels.

Motivated by these considerations, we introduce **VideoTapestry**, a training-free multi-agent framework that adapts a preconstructed hierarchical video memory through coarse-to-fine, query-driven refinement. The paradigm of our VideoTapestry is illustrated in Figure 1(c). The memory consists of SuperEvents, MacroEvents, and local Subgraphs, representing global stages, localized events, and fine-grained relational evidence. The Overview, Skim, and Focus agents operate at these levels, respectively, each within a scale-specific context. Guided by candidate memories and upstream inspection objectives, each agent acquires targeted multimodal evidence, refines a temporary layer memory, and decides whether to stop or pass selected branches and unresolved inspection objectives to the next level.

At the Super level, relevant descriptions and their neighboring context are retained and enriched, while the remaining events are summarized to preserve global coverage. Along selected branches, MacroEvent and Subgraph descriptions are refined with question-specific evidence. These layer-wise memories are assembled according to their original parent-child relations and paired with a concise query-time trace of inspection objectives and stopping decisions. The resulting **composite query-adaptive memory** preserves global context in a compact form while retaining detailed evidence along query-relevant branches. All refinements are query-specific and leave the preconstructed memory unchanged. An independent Answer Agent generates the final prediction from the assembled memory without further retrieval or observation.

We evaluate VideoTapestry on LVBench, LongVideoBench, Video-MME, and EgoSchema against representative direct VLM baselines and agentic approaches. Using GPT-5.5 as the base VLM, VideoTapestry improves accuracy over direct GPT-5.5 inference by 17.2, 14.9, 9.8, and 7.0 percentage points on LVBench, LongVideoBench (Long), Video-MME (Long), and EgoSchema, respectively. It achieves the best reported results among the compared methods across all four benchmarks.

Controlled ablation studies further demonstrate the effectiveness of query-adaptive memory refinement and scale-aligned multi-agent collaboration.

Our contributions are threefold:

- We propose a query-adaptive memory strategy that integrates newly observed evidence into selected branches while preserving global context and the original hierarchical structure.
- We develop **VideoTapestry**, a training-free multi-agent framework that aligns agents with memory levels and coordinates refinement through branch selection and inspection-objective transfer.
- We demonstrate consistent improvements across four video understanding benchmarks, with controlled ablations and cross-backbone evaluations supporting the effectiveness of memory refinement and scale-aligned collaboration.

2 Related Work

2.1 Long-Video Understanding

Long-video understanding extends the scope of VLM reasoning from short clips to complex real-world streams lasting many minutes or even hours. Recent work expands visual context windows, compresses redundant spatiotemporal tokens, and scales long-form instruction tuning through LongVU (Shen et al., 2024), LongViTU (Wu et al., 2025), and Hour-LLaVA (Lin et al., 2026b). Video Panels and Tempo further improve effective temporal coverage via frame grids and query-aware visual compression (Doorenbos et al., 2026; Fei et al., 2026). Despite these advances, practical context budgets still limit reliable evidence coverage, motivating agentic and memory-based methods that organize video content externally and acquire evidence through query-conditioned retrieval and observation (Fan et al., 2024; Zhang et al., 2026; Chen et al., 2026).

2.2 Hierarchical Video Memory

Video memory systems store frame descriptions, clips, or temporal segments outside the model context, while hierarchical and graph-based variants organize them across multiple semantic levels and explicit relations (Long et al., 2026; Yeo et al., 2026; Wang et al., 2025b; Zuo et al., 2026; Yin et al., 2026). PyraVid uses structure-guided expansion and pruning (Yan et al., 2026b), while VAM aligns temporal and spatial evidence (Li et al., 2026a). MAGIC-Video combines multimodal graph retrieval with narrative chains (Li et al., 2026b), while MERIT uses multi-key episodic retrieval with selective query-time filtering (Choi et al., 2026). VideoTapestry adapts a query-independent hierarchy through targeted re-observation, retaining a compact global scaffold and refined query-relevant branches in a composite query-adaptive memory.

2.3 Agentic and Multi-Agent Video Reasoning

Agentic video systems replace one-pass prediction with planning, temporal search, retrieval, and selective observation (Fan et al., 2024; Ye et al., 2025; Zhang et al., 2026; Lin et al., 2026a). Active Video Perception plans observations and checks evidence sufficiency (Wang et al., 2026b), while LensWalk controls temporal scope and sampling density (Li et al., 2026c). Route2Look routes queries among global browsing, temporal grounding, and semantic retrieval via a routing skill (Wang et al., 2026a). VideoXAgent acquires multimodal evidence online through expert tools (Yang et al., 2026). Multi-agent approaches distribute perception, localization, verification, and reasoning across specialized roles (Yan et al., 2026a; Rege et al., 2026; Zuo et al., 2026). VideoTapestry focuses on final hierarchical representation: scale-aligned agents retain a global scaffold, refine query-relevant branches, and assemble layer-wise memories for an independent Answer Agent.

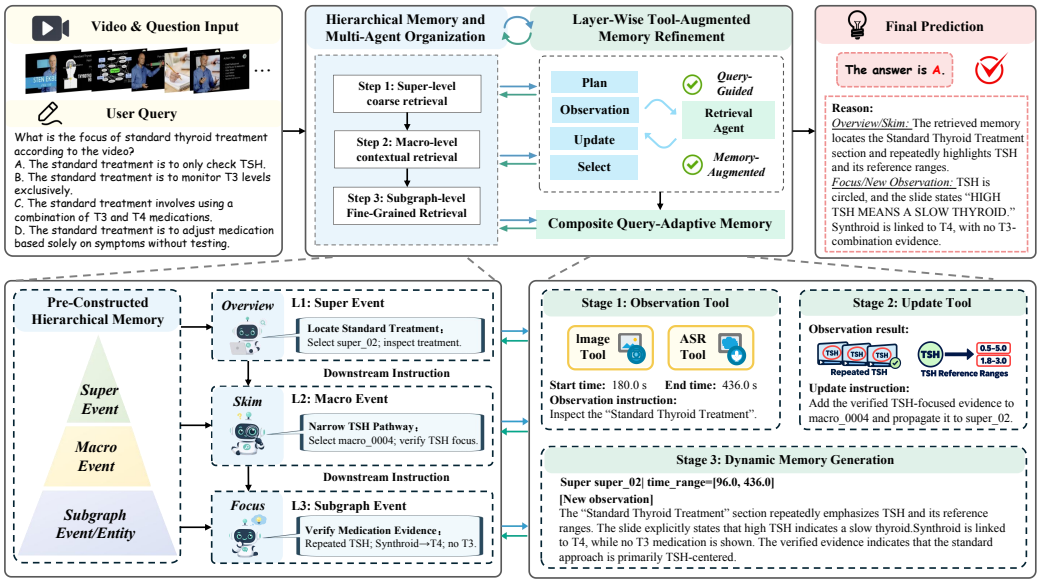


Figure 2: **Overview of VideoTapestry.** Scale-aligned agents operate over a preconstructed hierarchical video memory, while layer-wise tool-augmented refinement selectively incorporates newly acquired query-relevant observations at different semantic levels. The resulting refined layer memories are hierarchically assembled into a composite query-adaptive memory for final prediction.

3 METHOD

Given a long video V, a question q, and answer options o, VideoTapestry transforms a preconstructed hierarchical video memory into a composite query-adaptive memory. Three scale-aligned agents acquire evidence and refine temporary layer memories, which are then assembled for an independent Answer Agent (Figure 2).

3.1 QUERY-ADAPTIVE MEMORY AND SCALE-ALIGNED ORGANIZATION

Query-adaptive memory representation. VideoTapestry constructs a query-adaptive reasoning context by compressing unrelated SuperEvents, enriching relevant descriptions with new observations, and assembling refined descendants under their original parents. These operations modify only temporary textual representations while preserving the underlying static hierarchy.

We adopt the memory-construction procedure of MemDreamer (Chen et al., 2026) without modification. The resulting query-independent hierarchy $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ contains three semantic levels:

$$\mathcal{M}_{\text{static}} = \{\mathcal{M}^S_{\text{static}}, \mathcal{M}^M_{\text{static}}, \mathcal{M}^G_{\text{static}}\}.$$

SuperEvents summarize coarse temporal stages, MacroEvents describe localized coherent events, and each MacroEvent is associated with a complete local Subgraph containing Entity, Event, and OCRText nodes connected by typed relations. The levels S, M, and G provide global narrative context, event-level temporal structure, and fine-grained factual evidence, respectively.

Scale-aligned multi-agent organization. The Overview Agent refines the Super-level scaffold and selects relevant stages; the Skim Agent examines their MacroEvents and narrows candidate regions; and the Focus Agent inspects the corresponding Subgraphs for fine-grained verification. Each agent operates within a scale-specific context, with selected parent identifiers determining which descendants are exposed at the next level.

At level $\ell \in \{S, M, G\}$, the agent produces a temporary refined memory, selected identifiers, a downstream instruction, and a stop flag:

$$\left(\widetilde{\mathcal{M}}^\ell_q, \mathcal{S}^\ell, \mathcal{I}^\ell, z^\ell\right) = \mathcal{A}_\ell\left(q, o, s_V, \mathcal{M}^\ell_{\text{static}}[\mathcal{C}^\ell], \mathcal{I}^{\ell-1}, \mathcal{T}_\ell\right).$$

Here, s_V, $\mathcal{C}^\ell$, $\mathcal{I}^{\ell-1}$, and $\mathcal{T}_\ell$ denote the global video summary, candidate nodes, upstream instruction, and level-specific tools, respectively. If refinement continues, the selected identifiers determine the next candidate set:

$$\mathcal{C}^{\ell+1} = \bigcup_{v \in \mathcal{S}^\ell} \mathrm{Children}(v, \mathcal{G}).$$

The next agent loads these candidates from static memory. The instruction $\mathcal{I}^\ell$ conveys the inspection objective and may summarize intermediate findings and unresolved evidence needs without transmitting the full preceding-layer memory or tool trajectory.

The validated stop flag z^ℓ determines whether refinement terminates. In the evaluated Video-MME configuration, global questions may terminate at an intermediate level, whereas non-global questions must descend when valid descendants are available. The Focus Agent always terminates exploration. The question-type heuristic and decision validation rules are detailed in Appendix B.1.

3.2 Layer-Wise Tool-Augmented Memory Refinement

All agents follow the $\mathrm{Plan}_\ell \to \mathrm{Observe}_\ell \to \mathrm{Update}_\ell \to \mathrm{Decide}_\ell$ protocol. Plan identifies evidence needs, Observe inspects selected intervals, Update integrates new evidence into temporary memory, and Decide selects the next scope or terminates. The evaluated Video-MME configuration performs one refinement batch per visited layer, followed by one layer-level decision (Appendix B.1).

All levels access the source video through a Visual Observation Tool:

$$\mathcal{T}^\ell_{\mathrm{vis}}\left(\mathcal{W}^\ell_{\mathrm{vis}}, \mathcal{I}^\ell_{\mathrm{obs}}\right) \to \mathcal{O}^{\mathrm{vis},\ell}_q, \qquad \ell \in \{S, M, G\},$$

where $\mathcal{W}^\ell_{\mathrm{vis}}$ denotes a set of temporal intervals, $\mathcal{I}^\ell_{\mathrm{obs}}$ specifies what to inspect, and $\mathcal{O}^{\mathrm{vis},\ell}_q$ contains the resulting observations.

The Overview Agent additionally accesses a Transcript/ASR Tool:

$$\mathcal{T}_{\mathrm{txt}}\left(\mathcal{W}_{\mathrm{txt}}, \mathcal{I}_{\mathrm{txt}}\right) \to \mathcal{O}^{\mathrm{txt}}_q,$$

which extracts spoken-information evidence from timestamped text within selected windows. In the evaluated Video-MME configuration, it uses Whisper transcripts and follows visual observation on each successful Overview pass. The Skim and Focus Agents use visual observations only.

3.2.1 Overview Agent: Super-Level Memory Refinement

The Overview Agent accesses all SuperEvents. For local questions, it decomposes the localization objective into atomic targets. Entity, OCR-text, and semantic retrieval interfaces provided by the preconstructed memory (Chen et al., 2026) yield temporal anchors for observation; retrieval outputs serve as localization hints rather than verified answer evidence.

A visual planner converts these anchors into an inspection instruction and a continuous observation window bounded by the video duration. Sampled frames are arranged into grids for $\mathcal{T}^S_{\mathrm{vis}}$. In the evaluated Video-MME configuration, a subsequent ASR planner uses the SuperEvent timeline and visual results to select transcript windows and formulate a spoken-information request. The resulting observations are

$$\mathcal{O}^S_q = \mathcal{O}^{\mathrm{vis},S}_q \cup \mathcal{O}^{\mathrm{txt}}_q.$$

Observations are associated with SuperEvents using candidate identifiers and temporal overlap. Let $\mathcal{C}_{\mathrm{ev}}$ index the evidence-associated SuperEvents in temporal order. A protected set includes these events and their immediate neighbors:

$$\mathcal{P} = \left(\bigcup_{i \in \mathcal{C}_{\mathrm{ev}}} \{i-1, i, i+1\} \right) \cap \{1, \ldots, N_S\},$$

where N_S is the number of SuperEvents.

The Overview Agent updates the Super-level memory asymmetrically. Protected nodes retain their original descriptions and receive available associated evidence, while the remaining nodes are compressed:

$$\widetilde{m}^S_i = \begin{cases} m^S_i \oplus [\mathrm{Visual}]\, v_i \oplus [\mathrm{Transcript}]\, a_i \oplus [\mathrm{Evidence}]\, e_i, & i \in \mathcal{P}, \\ c(m^S_i), & i \notin \mathcal{P}. \end{cases}$$

Here, m_i^S is the original description, v_i contains visual observations, a_i contains transcript-derived facts, and e_i contains timestamped verbatim transcript evidence. The operator $\oplus$ denotes textual composition, and $c(\cdot)$ produces a compact summary. Unavailable evidence fields are omitted. The protected set is enforced programmatically: protected neighbors without associated observations retain their original text.

This refined global scaffold combines video-wide coverage with query-relevant detail. The agent then selects SuperEvents and a downstream instruction, or terminates according to the validated stopping rule.

3.2.2 Skim Agent: Macro-Level Memory Refinement

The Skim Agent operates on $\mathcal{C}^M = \bigcup_{v \in \mathcal{S}^S} \mathrm{Children}(v, \mathcal{G})$, the MacroEvent children of the selected SuperEvents.

Guided by the upstream instruction, it plans observations within candidate intervals and associates the resulting evidence with the covered nodes. For each requested update, it rewrites the original description:

$$\widetilde{\mathrm{text}}_i^M = \mathrm{Rewrite}_M\big(\mathrm{text}_i^M, q, o, \mathcal{I}^S, \mathcal{O}_i^M\big)\,.$$

The rewrite preserves relevant static facts, incorporates observed evidence, and compresses unrelated content. Missing or invalid updates retain the original text. Node identifiers, parent relations, semantic levels, and temporal ranges remain fixed. The agent then selects MacroEvents and generates an instruction for finer inspection, subject to the stopping rule.

3.2.3 Focus Agent: Subgraph-Level Memory Refinement

The Focus Agent loads the complete local Subgraphs associated with the selected MacroEvents, exposing entities, events, OCR content, and their typed relations. Guided by the upstream instruction, it plans fine-grained visual inspection and rewrites the serialized description of each requested Subgraph:

$$\widetilde{\mathrm{text}}_i^G = \mathrm{Rewrite}_G\big(\mathrm{text}_i^G, q, o, \mathcal{I}^M, \mathcal{O}_i^G\big)\,.$$

This operation modifies the temporary textual description rather than the underlying graph nodes or edges. As at the Macro level, missing or invalid updates retain the original text. The Focus Agent then terminates hierarchical exploration.

3.3 Composite Memory Assembly and Answer Generation

After traversal, $\mathrm{Assemble}$ retains the complete refined Super-level scaffold and places selected refined MacroEvent and Subgraph memories beneath their original parents. This preserves video-wide event coverage while allocating fine-grained content to query-relevant branches.

The assembled memory is serialized together with the question, options, and a compact handoff trace:

$$\mathcal{D}_q = \mathrm{Serialize}\Big(q, o, H_q, \mathrm{Assemble}\Big(\big\{\widetilde{\mathcal{M}}_q^\ell\big\}_{\ell \in \mathcal{L}_q}\Big)\Big)\,.$$

Here, $\mathcal{L}_q \subseteq \{S, M, G\}$ contains the visited levels, and $H_q = [(a_\ell, i_\ell)]_{\ell \in \mathcal{L}_q}$ pairs each agent identifier with its handoff instruction or stopping marker. Instructions may summarize intermediate findings and unresolved evidence needs; the trace does not include complete observation or update logs.

The independent Answer Agent predicts

$$\widehat{y} = \mathcal{A}_{\mathrm{ans}}(q, o, \mathcal{D}_q)$$

without further retrieval or observation.

Table 1: Comparison with representative VLMs and agentic approaches on four long-video understanding benchmarks. Results are reported as accuracy (%). Green parenthesized values denote absolute percentage-point gains over GPT-5.5 (Direct). † denotes evaluation with the frozen GPT-5.5 backbone, and ‡ denotes results from our reproduction.

Method	Query-time Model(s)	LVBench	LongVideoBench		Video-MME(w/o)	EgoSchema
		Overall	Val	Long	Long	Val
Direct Proprietary VLMs						
Gemini-2.5-Pro	Gemini-2.5-Pro	72.0	71.0	68.6	75.9	72.8
Gemini-3.1-Pro	Gemini-3.1-Pro	78.2	78.6	77.0	80.3	76.4
OpenAI o3	OpenAI o3	57.1	66.7	60.6	64.7	63.2
Direct Open-Source VLMs						
Seed1.5VL	Seed1.5VL-Thinking	64.6	–	74.4	–	–
Qwen3-VL	Qwen3-VL-235B-A22B-Thinking	63.6	71.4	67.1	71.2	75.9
Qwen3.5	Qwen3.5-35B-A3B	71.3	–	–	62.6	–
GLM-4.6V	GLM-4.6V-106B-A12B	59.5	67.6	58.7	66.0	68.8
Agentic Methods						
DVD	OpenAI o3	74.2	71.6	68.6	67.3	76.6
Symphony	DeepSeek-R1 + Seed1.6VL	71.8	77.1	–	78.1	–
VideoARM	o3 + GPT-4.1	79.7	73.7	69.2	75.3	78.2
VideoARM	o3 + GPT-4o	–	78.0	76.4	81.2	76.2
Active Video Perception	Gemini-2.5-Pro	74.8	73.4	70.0	81.9	–
WorldMM-GPT	OpenAI GPT-5	61.9	–	–	76.6	–
VideoSeek	OpenAI GPT-5	68.4	–	73.5	70.1	–
LensWalk	OpenAI GPT-5	66.9	–	68.8	69.2	74.6
Route2Look	OpenAI GPT-5	75.4	–	77.8	76.1	–
VideoXAgent	Claude-Opus-4.6 + Gemini-3.1-Pro-preview	76.5	–	78.1	–	–
MemDreamer‡	OpenAI GPT-5.5	73.3	74.1	73.9	77.1	78.6
Controlled Comparison						
GPT-5.5 (Direct)†	OpenAI GPT-5.5	64.3	69.8	64.0	74.8	76.0
VideoTapestry†	**OpenAI GPT-5.5**	**81.5** (+17.2)	**80.4** (+10.6)	**78.9** (+14.9)	**84.6** (+9.8)	**83.0** (+7.0)

4 Experiments

4.1 Benchmarks

We evaluate VideoTapestry on four long-video understanding benchmarks. LVBench comprises 103 videos ranging from 30 minutes to two hours and 1,549 QA pairs across six fine-grained dimensions (Wang et al., 2025a). The LongVideoBench validation set contains 1,337 questions from 752 videos, including 564 questions from 188 long videos ranging from 15 to 60 minutes (Wu et al., 2024). Video-MME (Long) consists of 300 videos ranging from 30 to 60 minutes and 900 QA pairs (Fu et al., 2025). EgoSchema contains egocentric clips designed to evaluate video reasoning; for efficient and representative evaluation, we use a 500-question subset (Mangalam et al., 2023).

4.2 Experimental Setup

VideoTapestry is training-free, with all model parameters and the preconstructed hierarchical memory fixed during evaluation. Unless otherwise stated, GPT-5.5 (OpenAI, 2026) is used for the Overview, Skim, Focus, and Answer agents, while Whisper (Radford et al., 2023) is used for ASR. Qwen3-Embedding (Zhang et al., 2025) supports semantic retrieval during the search stage. The Direct baseline uses the same frozen backbone and evaluation examples without the agent framework.

Table 2: Comparison with representative VLMs and agentic approaches across the six capability dimensions of LVBench (accuracy, %). The best results are bolded; green parenthesized values denote absolute percentage-point gains over GPT-5.5 (Direct). ‡ denotes results from our reproduction.

Method	ER	EU	Rea	KIR	Sum	TG	Avg.
Direct VLMs							
GPT-4o	48.9	49.5	50.3	48.1	50.0	40.9	48.9
OpenAI o3	57.6	56.4	50.8	62.9	67.2	46.8	57.1
GPT-5.5 (Direct)	63.2	66.0	72.7	63.8	65.4	43.1	64.3
Structured and Agentic Methods							
VideoTree	30.3	25.1	31.9	26.5	25.5	27.7	28.8
VideoLucy	54.3	59.8	55.9	75.6	49.1	51.7	58.8
DVD	73.4	73.3	70.7	80.4	74.1	72.3	74.2
DVD + Transcripts	75.5	77.1	68.7	79.0	84.5	72.7	76.0
Symphony	70.0	69.4	69.4	77.2	72.5	70.1	71.8
Active Video Perception	71.9	76.7	67.7	80.1	75.9	73.6	74.8
MemDreamer (GPT-5.5)‡	70.5	74.9	73.5	80.1	67.3	67.7	73.3
VideoTapestry (GPT-5.5)	**80.7** (+17.5)	**81.0** (+15.0)	**77.6** (+4.9)	**84.5** (+20.7)	**89.7** (+24.3)	**83.2** (+40.1)	**81.5** (+17.2)

Following VideoARM (Yin et al., 2026), all ablations are conducted on a fixed 200-question subset of Video-MME Long.

Baselines. We compare with direct VLMs and agentic methods, including DVD (Zhang et al., 2026), Symphony (Yan et al., 2026a), VideoARM (Yin et al., 2026), Active Video Perception (Wang et al., 2026b), WorldMM (Yeo et al., 2026), VideoSeek (Lin et al., 2026a), LensWalk (Li et al., 2026c), Route2Look (Wang et al., 2026a), VideoXAgent (Yang et al., 2026), and MemDreamer (Chen et al., 2026).

4.3 Main Results

With frozen GPT-5.5, VideoTapestry gains 7.0–17.2 percentage points over Direct and surpasses the strongest prior result in each column by 0.8–4.4 points (Table 1). It also leads across all six LVBench dimensions, with the largest same-backbone gains in temporal grounding (+40.1) and summarization (+24.3) (Table 2). These results suggest benefits for both long-range evidence localization and global information aggregation.

The LVBench breakdown in Table 2 shows that VideoTapestry leads all six capability dimensions, improving over the strongest prior result by 3.9–9.6 points. The largest gains occur in temporal grounding (+9.6), supporting the effectiveness of coarse-to-fine retrieval and targeted observation for locating and verifying question-relevant evidence.

4.4 Ablation Studies

Following VideoARM (Yin et al., 2026), all controlled studies use GPT-5.5 on a 200-question subset of Video-MME Long, with the same preconstructed memory and query-time inference budget.

Query-time memory refinement. Table 3(a) examines query-specific evidence acquisition and integration into hierarchical memory. The two observation ablations remove visual observation or selected-window ASR. *Append-only Evidence* retains the same tools and agent structure but provides new evidence as a flat log without updating selected memory nodes. *Single Agent* replaces the three scale-aligned agents with one controller under the same model- and tool-call budgets. Removing visual observation, ASR, structured memory refinement, and multi-agent collaboration reduces accuracy by 3.0, 5.5, 1.5, and 4.5 points, respectively. The largest drops come from removing ASR and scale-aligned collaboration, while structured memory refinement outperforms flat evidence logging.

Table 3: Ablations of (a) query-time memory refinement and (b) memory configurations. Update denotes structured evidence integration; M-A denotes scale-aligned multi-agent collaboration. Δ Acc. is the percentage-point difference from VideoTapestry. In (b), Super and Paths indicate refined components; – denotes the unmodified static-memory baseline.

(a) Query-time memory refinement

Variant	Visual	ASR	Update	M-A	Acc.	Δ Acc.
w/o Visual Observation	N	Y	Y	Y	84.0	−3.0
w/o ASR Observation	Y	N	Y	Y	81.5	−5.5
Append-only Evidence	Y	Y	N	Y	85.5	−1.5
Single Agent	Y	Y	Y	N	82.5	−4.5
VideoTapestry	Y	Y	Y	Y	**87.0**	–

(b) Memory configurations

Memory Variant	Super	Paths	Tokens	Acc.
Static Memory Only	–	–	4,874	79.5
Refined Super Only	Y	N	3,250	85.0
Refined Paths Only	N	Y	1,363	80.5
VideoTapestry	Y	Y	4,417	**87.0**

Composite memory. Table 3(b) compares variants of the composite query-adaptive memory. *Static Memory Only* retrieves the preconstructed memory without query-time updates, compression, or rewriting, achieving 79.5%. The refined variants share the same retrieval results, observations, and query-time trace, but differ in the video memory provided to the Answer Agent: *Refined Super Only* combines compressed irrelevant SuperEvents with preserved relevant descriptions and new observations (85.0%), while *Refined Paths Only* retains refined Macro- and Subgraph-level memories without the Super-level memory (80.5%). Combining both components reaches 87.0%, improving over static retrieval by 7.5 percentage points and outperforming either refined component alone, supporting their complementary roles.

4.5 Backbone Generalization

Table 4 evaluates backbone generalization while keeping the memory and tool interfaces fixed. We replace the backbone used by the agents and Answer Agent with GPT-5.5, GPT-5, Qwen3.7-Plus, and Qwen3.6-27B (OpenAI, 2025; Alibaba Cloud, 2026; Qwen Team, 2026). Direct and VideoTapestry use the same backbone and evaluation examples. VideoTapestry improves every backbone by 7.0–9.5%, with the largest gain of 9.5% on Qwen3.6-27B. These consistent gains across GPT and Qwen show the method generalizes beyond a single model family.

Table 4: Generalization across Agent and Answer backbones.

Backbone	Direct	VideoTapestry	Improve.
GPT-5.5	78.0	87.0	+9.0
GPT-5	75.5	82.5	+7.0
Qwen3.7-Plus	75.0	83.0	+8.0
Qwen3.6-27B	65.5	75.0	+9.5

5 Conclusion

We introduce VideoTapestry, a training-free hierarchical multi-agent framework for long-video understanding. Starting from a query-independent hierarchical video memory, the Overview, Skim, and Focus agents perform coarse-to-fine retrieval and targeted observation, refining query-relevant content at complementary semantic granularities. Their layer-wise refinements are assembled with the global scaffold into a composite query-adaptive memory that preserves video-wide structure while enriching relevant branches for final reasoning. Experiments on long-video benchmarks show that VideoTapestry consistently improves over the same-backbone direct baseline and representative agentic methods, demonstrating the effectiveness of organizing global structure and query-specific evidence for final reasoning.

AI USE STATEMENT

Generative AI tools were used to assist with literature search, methodological brainstorming, experimental planning, and the drafting and polishing of portions of the manuscript. All AI-assisted suggestions and content were independently reviewed and verified by the authors, including suggested citations and technical descriptions. All reported experimental results were obtained and verified by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Closed- and open-source baseline evaluation. Values in the main table are obtained from the corresponding models' technical reports or previously published benchmark results when available; the remaining entries are obtained from our evaluations. GPT- and Qwen-based models are evaluated via their corresponding APIs across all four benchmarks. Direct baselines receive uniformly sampled video frames. Under the maximum image-upload limits of the API gateways used in our experiments, GPT-family models receive 50 frames, Qwen3.7-Plus receives 196 frames, and Qwen3.6-27B receives 396 frames per video. For GPT-family models, the reasoning effort is set to high to enable extended reasoning, while Qwen-family models use thinking mode during direct inference. All inference runs use a temperature of 0.1.

VideoTapestry inference configuration and memory construction. Each Visual Observation Tool call uploads at most 50 images, respecting the GPT API gateway limit. Overview samples up to 120 frames and packs them into at most 20 six-frame grids, whereas Skim and Focus each use up to 50 individual frames from selected intervals. For fair comparison, the ASR Tool applies Whisper only to agent-selected audio windows rather than transcribing the full video. For GPT-family models, only the Answer Agent uses high reasoning effort, while all other calls use the default setting. For Qwen-based experiments, thinking is enabled only for planning, decision, and final answering, while all other calls use non-thinking mode. To isolate our query-time method from offline memory construction, all experiments use hierarchical video memories built following the MemDreamer procedure (Chen et al., 2026). For the MemDreamer baseline, we separately reproduce its retrieval pipeline.

REFERENCES

Alibaba Cloud. Qwen3.7-Plus model info. Alibaba Cloud Model Studio Documentation, June 2026. URL https://www.alibabacloud.com/help/en/model-studio/qwen3-7-plus.

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*, 2026.

Yeeun Choi, Youngbeom Yoo, Joon-Young Lee, Hyolim Kang, and Seon Joo Kim. Keep it simple: Multi-key episodic memory retrieval for ultra-long video understanding. *arXiv preprint arXiv:2608.07663*, 2026.

Lars Doorenbos, Federico Spurio, and Juergen Gall. Video panels for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31293–31303, 2026.

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.

Junjie Fei, Jun Chen, Zechun Liu, Yunyang Xiong, Chong Zhou, Wei Wen, Junlin Han, Mingchen Zhuge, Saksham Suri, Qi Qian, et al. Small vision-language models are smart compressors for long video understanding. *arXiv preprint arXiv:2604.08120*, 2026.

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

Aiden Yiliu Li, Nels Numan, and Anthony Steed. Visual agentic memory: Enabling online long video understanding via online indexing, hierarchical memory, and agentic retrieval. *arXiv preprint arXiv:2605.16481*, 2026a.

Jiazheng Li, Chi-Hao Wu, Yunze Liu, Kaize Ding, Jundong Li, and Chuxu Zhang. Bridging modalities, spanning time: Structured memory for ultra-long agentic video reasoning. *arXiv preprint arXiv:2605.08271*, 2026b.

Keliang Li, Yansong Li, Hongze Shen, Mengdi Liu, Hong Chang, and Shiguang Shan. Lenswalk: Agentic video understanding by planning how you see in videos. *arXiv preprint arXiv:2603.24558*, 2026c.

Jingyang Lin, Jialian Wu, Jiang Liu, Ximeng Sun, Ze Wang, Xiaodong Yu, Jiebo Luo, Zicheng Liu, and Emad Barsoum. Videoseek: Long-horizon video agent with tool-guided seeking. *arXiv preprint arXiv:2603.20185*, 2026a.

Jingyang Lin, Jialian Wu, Ximeng Sun, Ze Wang, Jiang Liu, Yusheng Su, Xiaodong Yu, Hao Chen, Jiebo Luo, Zicheng Liu, et al. Unleashing hour-scale video training for long video-language understanding. *Advances in Neural Information Processing Systems*, 38:17523–17552, 2026b.

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 *International Conference on Learning Representations*, volume 2026, pp. 146197–146246, 2026.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

OpenAI. GPT-5 system card, August 2025. URL `https://openai.com/index/gpt-5-system-card/`.

OpenAI. GPT-5.5 system card, April 2026. URL `https://deploymentsafety.openai.com/gpt-5-5`.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026. URL `https://qwen.ai/blog?id=qwen3.6-27b`.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Aniket Rege, Arka Sadhu, Yuliang Li, Kejie Li, Ramya Korlakai Vinayak, Yuning Chai, Yong Jae Lee, and Hyo Jin Kim. Agentic very long video understanding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46575–46602, 2026.

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.

Tianyue Wang, Xuying Wu, Yuxiang Ma, Ruiming Liang, Jiaxuan Kang, Yanchao Hao, Zheng Wei, Leigang Qu, Haiyun Guo, and Jinqiao Wang. Routing before looking: Query-adaptive evidence acquisition for long-form video understanding. *arXiv preprint arXiv:2608.20805*, 2026a.

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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22958–22967. IEEE, 2025a.

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, 2025b.

Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S Ryoo, and Juan Carlos Niebles. Active video perception: Iterative evidence seeking for agentic long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9088–9099, 2026b.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Rujie Wu, Xiaojian Ma, Hai Ci, Yue Fan, Yuxuan Wang, Haozhe Zhao, Qing Li, and Yizhou Wang. Longvitu: Instruction tuning for long-form video understanding. *arXiv preprint arXiv:2501.05037*, 2025.

Haiyang Yan, Hongyun Zhou, Peng Xu, Xiaoxue Feng, and Mengyi Liu. Symphony: A cognitively-inspired multi-agent system for long-video understanding. *arXiv preprint arXiv:2603.17307*, 2026a.

Sikuan Yan, Sicheng Dong, Haotong Wang, Ercong Nie, Yilun Liu, Jinhe Bi, Yingjie Xu, Susanna Schwarzmann, Riccardo Trivisonno, Volker Tresp, et al. Pyravid: Hierarchical multimodal memory for long-horizon video reasoning. *arXiv preprint arXiv:2605.17065*, 2026b.

Sen Yang, Boqiang Duan, Jing Yang, Weihao Bo, Jie Liu, Boyuan Tong, Ze Feng, Wenkang Zhang, Jingdong Wang, and Hua Wu. Online video agent harness for long video understanding. *arXiv preprint arXiv:2609.12818*, 2026.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, et al. Re-thinking temporal search for long-form video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8579–8591. IEEE, 2025.

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.

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.

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

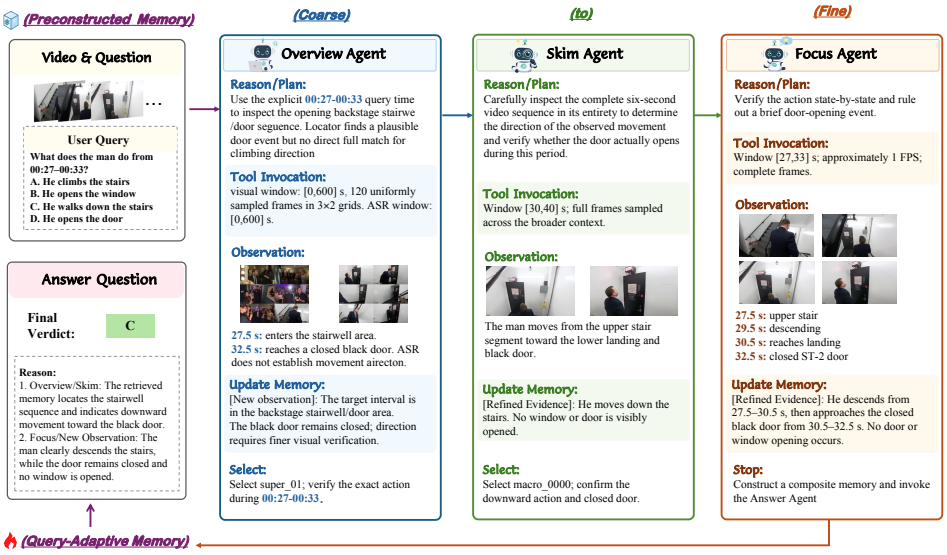
Jialong Zuo, Yongtai Deng, Lingdong Kong, Jingkang Yang, Rui Jin, Yiwei Zhang, Nong Sang, Liang Pan, Ziwei Liu, and Changxin Gao. Videolucy: Deep memory backtracking for long video understanding. *Advances in Neural Information Processing Systems*, 38:25660–25691, 2026.

A ADDITIONAL RESULTS

Additional Qualitative Results. Figure 3 shows two representative trajectories from coarse localization to fine-grained verification, illustrating progressive query-adaptive memory refinement.

Case 1: Fine-grained temporal verification.

The agents narrow the stairwell interval to 27–33 s, verify descent without door or window opening, and answer C.



Case 2: Cross-branch evidence aggregation.

The agents verify distinct performers across two piano-related branches roughly 25 minutes apart, aggregate the evidence, and answer D.

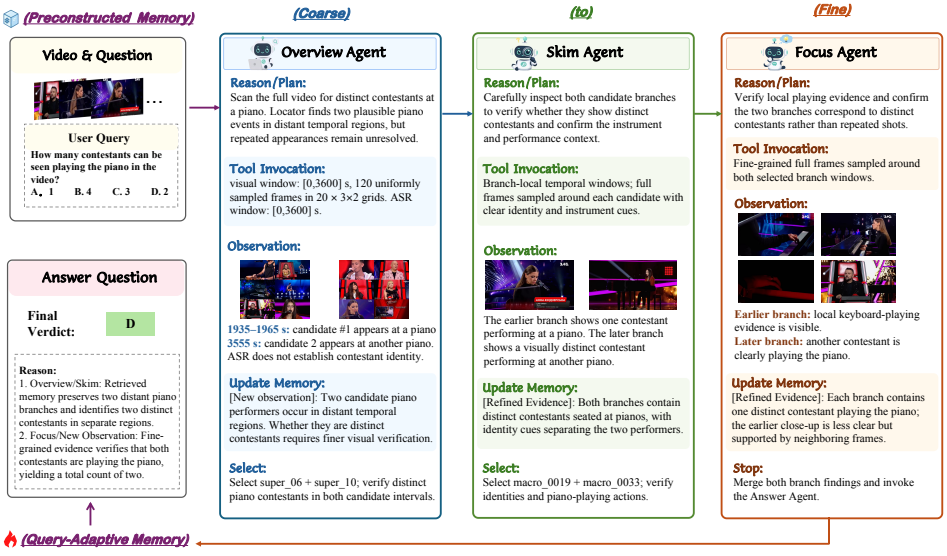


Figure 3: Two examples of coarse-to-fine query-adaptive memory refinement.

B INFERENCE DETAILS

B.1 DETAILED INFERENCE PIPELINE

Algorithm 1 summarizes hierarchical exploration and final memory assembly. Algorithm 2 specifies layer execution in the evaluated Video-MME configuration: each visited layer processes one batch, followed by one decision. Overview performs visual observation and then Whisper-based ASR; Skim and Focus use visual observation. There is no repeated within-layer refinement in this configuration.

Algorithm 1: Inference Pipeline of VideoTapestry

Input: Video V, question q, options o, static hierarchy $\mathcal{G}$ and memory $\mathcal{M}_{\text{static}}$, refinement agents $\{\mathcal{A}_S, \mathcal{A}_M, \mathcal{A}_G\}$, and Answer Agent $\mathcal{A}_{\text{ans}}$.
Output: Predicted answer $\widehat{y}$.

Initialize $\mathcal{C}^S$ with all SuperEvents, $\mathcal{I}^0 \leftarrow \varnothing$, $\mathcal{L}_q \leftarrow \varnothing$, and trace $H_q \leftarrow [\,]$.
for $\ell \in \{S, M, G\}$, aligned with Overview, Skim, and Focus, **do**
  Load original candidate memory $\mathcal{M}^\ell \leftarrow \mathcal{M}^\ell_{\text{static}}[\mathcal{C}^\ell]$.
  if $\mathcal{M}^\ell = \varnothing$, **break**.
  Run Algorithm 2 to obtain $(\widetilde{\mathcal{M}}^\ell_q, \mathcal{S}^\ell, \mathcal{I}^\ell, z^\ell)$.
  $\mathcal{L}_q \leftarrow \mathcal{L}_q \cup \{\ell\}$; append the agent's handoff instruction or stopping marker to H_q.
  if z^ℓ is true, **break**.
  $\mathcal{C}^{\ell+1} \leftarrow \bigcup_{v \in \mathcal{S}^\ell} \text{Children}(v, \mathcal{G})$.
end for
$\mathcal{D}_q \leftarrow \text{Serialize}\Big(q, o, H_q, \text{Assemble}\Big(\{\widetilde{\mathcal{M}}^\ell_q\}_{\ell \in \mathcal{L}_q}\Big)\Big)$.
Return $\widehat{y} \leftarrow \mathcal{A}_{\text{ans}}(q, o, \mathcal{D}_q)$ without further retrieval or observation.

Algorithm 2: Layer-Wise Observation and Memory Refinement

Input: Level ℓ, video V, question q, options o, nonempty static candidate memory $\mathcal{M}^\ell$, and upstream instruction $\mathcal{I}^{\ell-1}$.
Output: Refined memory $\widetilde{\mathcal{M}}^\ell_q$, selected IDs $\mathcal{S}^\ell$, downstream instruction $\mathcal{I}^\ell$, and stop flag z^ℓ.

Initialize local trajectory $H_\ell \leftarrow [\,]$ and observations $O \leftarrow \varnothing$; $g \leftarrow \text{GlobalHeuristic}(q, o)$.
$p \leftarrow \text{Plan}_\ell(q, o, \mathcal{M}^\ell, \mathcal{I}^{\ell-1}, H_\ell)$; at S, include locator hints.
if $\ell = S$ **then**
  Merge planned ranges into one video-bounded window; fall back to the full video if invalid, and expand to at least $\min(600\,\text{s}, \text{duration}(V))$.
  $O_v \leftarrow \text{VisualObserve}$ using up to 120 frames packed into 20 six-frame grids.
  $p_a \leftarrow \text{PlanASR}(q, o, \mathcal{M}^S, p, O_v)$; fall back to the visual window if no ASR ranges are proposed.
  Normalize up to three ASR windows to 600–1200 s each within video bounds (use the full video if shorter than 600 s).
  $O_a \leftarrow \text{ASRObserve}(p_a)$ using Whisper transcripts; $O \leftarrow \{O_v, O_a\}$.
  Construct $\mathcal{P}$ from evidence-linked SuperEvents and their valid immediate temporal neighbors.
  Build $\widetilde{\mathcal{M}}^S_q$: retain original text in $\mathcal{P}$ and append available associated evidence; compress nodes outside $\mathcal{P}$.
else
  if $p.\text{observation_needed}$ or $(\ell = M$ and $\neg g)$ **then**
    Intersect planned ranges with candidate intervals (fallback: candidate intervals); $O \leftarrow$ VisualObserve using up to 50 individual frames.
  end if; $\widetilde{\mathcal{M}}^\ell_q \leftarrow \text{Update}_\ell(q, o, \mathcal{I}^{\ell-1}, \mathcal{M}^\ell, p, O)$ for valid requested nodes.
end if; record plans, observations, and updates in H_ℓ.
$(\mathcal{S}^\ell, \mathcal{I}^\ell, z^\ell) \leftarrow \text{Decide}(q, o, \mathcal{M}^\ell, \widetilde{\mathcal{M}}^\ell_q, \mathcal{I}^{\ell-1}, H_\ell)$ once.
Force $z^\ell \leftarrow \text{true}$ at G; at S or M, force $z^\ell \leftarrow$ false for non-global questions. Validate selections and apply fallback IDs/instructions when continuing; stop if no valid descendants remain.
If stopping, clear $\mathcal{S}^\ell$ and $\mathcal{I}^\ell$; return $(\widetilde{\mathcal{M}}^\ell_q, \mathcal{S}^\ell, \mathcal{I}^\ell, z^\ell)$.

B.2 Prompts Used in VideoTapestry

We provide the core prompts used throughout our framework, covering the main stages of the overall pipeline and the key components involved in its execution, including planning, observation, memory updating, decision making, and answer generation.

B.2.1 Hierarchical Observation Planning

Overview Agent Planning Prompt

```
YOU ARE OVERVIEW AGENT.
You are responsible for updating the current-level hierarchical video
     memory under a specific question.
You can inspect visual frames through OBSERVATION TOOL and rewrite
    old memory text through UPDATE TOOL.
Your goal is to produce concise, accurate, query-relevant dynamic
    memory for the current layer.
Old memory text is only a static prebuilt summary. It is not visual
    verification and must not be treated as a substitute for
    observation. Only set observation_needed=false when the video
    file or visual frames are unavailable, or when this layer already
     has a prior visual observation in trajectory.
For Overview/Super, observation_needed must be true for the first
    update attempt.
Decide one bounded observation/update batch. Do not answer the
    question.
Return strict JSON with keys: thought, observation_needed,
    observation_instruction, candidate_memory_ids, time_ranges,
    locator_time_hints_used.
For Overview/Super, return exactly one contiguous visual observation
    time range inside observation_time_bounds. Center it on the
    evidence region you need to inspect; it may extend beyond an
    individual Super boundary. Choose a span of at least 600 seconds
    when the video duration permits; the program enforces the final
    minimum span.
overview_locator_results, when present, are only global time-
    localization leads from entity/OCR/node search. Do not treat
    their text as visual facts or as support for an option. Use their
     timestamps only to choose a question-relevant continuous
    observation window, then require Observation Tool to verify the
    fact visually.
When tool=multi_locator_compressed, candidate_windows are complete
    joint hypotheses containing one grounded match per atomic anchor.
     Prefer a complete, exact, temporally compact cluster; do not
    substitute an earlier partial look-alike merely because it ranks
    highly. Expand the chosen joint time_range into the required
    continuous observation span. If candidate_windows is empty or
    unresolved_anchors is nonempty, do not narrow the observation
    from partial Locator hits; instead use the complete Super
    memories to choose a broader window. Preserve plausible competing
     complete windows in locator_time_hints_used. The compressed
    result remains unverified localization, not answer evidence.
Return locator_time_hints_used as a list of the temporal hints you
    actually adopted. Each item must explain only why a time window
    was selected; do not copy raw search output, query text, option
    language, or unverified factual claims. Return [] if none was
    used.
candidate_memory_ids must come from supplied memory_id values.
Current level is super. Respect the tool limits: at most 2
    observations, at most 2 updates, at most 4 total calls.
```

Skim / Focus Planning Template

```
YOU ARE {AGENT_NAME}.
You are responsible for updating the current-level hierarchical video
     memory under a specific question.
You can inspect visual frames through OBSERVATION TOOL and rewrite
    old memory text through UPDATE TOOL.
Your goal is to produce concise, accurate, query-relevant dynamic
    memory for the current layer.
Old memory text is only a static prebuilt summary. It is not visual
    verification and must not be treated as a substitute for
    observation. Only set observation_needed=false when the video
    file or visual frames are unavailable, or when this layer already
     has a prior visual observation in trajectory.
For Overview/Super, observation_needed must be true for the first
    update attempt.
Decide one bounded observation/update batch. Do not answer the
    question.
Return strict JSON with keys: thought, observation_needed,
    observation_instruction, candidate_memory_ids, time_ranges,
    update_memory_ids.
time_ranges must stay inside the time_range of the supplied
    old_memories.
candidate_memory_ids and update_memory_ids must come from supplied
    memory_id values.
Current level is {MEMORY_LEVEL}. Respect the tool limits: at most 2
    observations, at most 2 updates, at most 4 total calls.
```

B.2.2 VISUAL OBSERVATION TOOL

Visual Observation Tool Prompt

```
YOU ARE OBSERVATION TOOL.
You are being called by {AGENT_NAME} for the {MEMORY_LEVEL} layer.
Inspect the provided visual frames and answer only the supplied
    observation_instruction. Do not update memory, infer an unstated
    question, or choose a multiple-choice answer.
For observation_result: describe specific visible facts (people,
    actions, objects, on-screen text, timestamps shown). Include only
     facts that directly address observation_instruction.
For uncertainty: list only facts you are unsure about.
For relevance_to_question: one sentence summary of how the frames
    answer observation_instruction.
If a fact is unclear, write it as uncertain instead of a confirmed
    fact.
Return strict JSON with keys: observation_result, uncertainty,
    relevance_to_question.
```

ASR Observation Tool Prompt

```
YOU ARE THE OVERVIEW AGENT'S ASR OBSERVATION TOOL.

You will receive exactly two inputs:
1. observation_instruction: the self-contained spoken-information
    request;
2. asr: the timestamped ASR transcript from the selected video window

You do not receive the original question or answer options. Analyze
    only the
supplied ASR to answer observation_instruction with objective facts.

OUTPUT REQUIREMENTS

1. observation_result
- Answer observation_instruction using only facts supported by the
    ASR.
- When evidence is conclusive, state the underlying fact clearly.
- When the evidence is incomplete, conflicting, or unclear,
    explicitly describe
  what cannot be determined.
- Do not use outside knowledge and do not select a multiple-choice
    answer.
2. asr_evidence
- Copy the most important supporting ASR lines exactly.
- Preserve timestamps, text, punctuation, capitalization, line breaks
    , and
  transcription errors.
- Do not paraphrase, translate, summarize, correct, merge, reorder,
    or rewrite.
- Include only directly supporting lines, in chronological order.
- Every character must come from the supplied asr.
- Return an empty string when there is no reliable evidence.

Return strict JSON with exactly these two fields:

{
  "observation_result": "...",
  "asr_evidence": "..."
}
```

B.2.3 MEMORY UPDATE TOOL

SuperEvent Update Prompt

```
YOU ARE UPDATE TOOL.

You are being called by Overview Agent for the super layer.

You receive question, options, previous_instruct, one shared
    update_instruction,
observation_sources, and update_requests. Each long observation
    source appears
exactly once in observation_sources. Every update request contains
    memory_id,
old_memory_text, and source_refs. Resolve source_refs against
    observation_sources.
Different Super memories may reference the same source so they can
    extract
different, time-local or narrative-context facts without duplicating
    the source
text in the request. Do not merge visual and ASR observations.

Do not modify memory_id, time_range, level, or parent_id.

SUPER LAYER RULES:

- First set is_relevant to true or false.

- If is_relevant is true:
  1. Write only concise objective visual facts from referenced visual
     sources
     into visual_observation. If no visual source is referenced,
    return empty.
  2. Write only concise objective spoken facts from referenced ASR
    sources into
     asr_observation. If no ASR source is referenced, return empty.
  3. Keep the two modalities separate. Do not move or duplicate facts
     between them.
  4. Do not repeat, summarize, rewrite, or quote old_memory_text.
  5. Do not return or quote raw asr_evidence. The program assigns
    original ASR
     lines to memories from their timestamps after this call.
  6. Keep irrelevant_summary empty.

- If is_relevant is false:
  1. Keep visual_observation and asr_observation empty.
  2. Write a 1-2 sentence objective summary of the segment's topic
    and
     key content in irrelevant_summary.
  3. Do not explain why the segment is irrelevant.

Do not include a "[Visual observation]", "[ASR observation]", or
"[ASR evidence]" heading. The program adds these headings when
    writing memory.

STRICT PROHIBITION:

Never write any option-tendency judgment in memory text.
Never write phrases such as "consistent with B", "supports option A",
"more likely X", "the correct answer is", or any equivalent statement
that steers toward a specific answer choice.

Only state objective facts supported by old memory or the supplied
```

```
observation. The final answer agent will judge the options.

Do not invent information that is not supported by old_memory_text or
     referenced
observation sources.

Preserve every supplied memory_id exactly.

Return strict JSON exactly in this schema:

{
  "updated_memories": {
    "memory_id": {
      "is_relevant": true,
      "visual_observation": "...",
      "asr_observation": "...",
      "irrelevant_summary": ""
    }
  }
}
```

MacroEvent / Subgraph Update Prompt

```
YOU ARE UPDATE TOOL.
You are being called by {AGENT_NAME} for the {MEMORY_LEVEL} layer.
Read the old memory text, question, options, and trajectory
    observations. Do not modify memory_id, time_range, level, or
    parent_id.
{MEMORY_LEVEL_UPPERCASE} LAYER RULES:
- RELEVANT memory: preserve specific facts - visible people, actions,
     objects, on-screen text, timestamps. Do not replace concrete
    details with vague phrases.
- IRRELEVANT memory: write an extremely short summary (<=10
    characters).
STRICT PROHIBITION: Never write any option-tendency judgment in text.
     Never write phrases like 'consistent with B', 'supports option A
    ', 'more likely X', or any statement that steers toward a
    specific answer choice. Only state observed objective facts. The
    final answer agent will judge the options.
Do not invent information not supported by old memory or observations
    .
Return strict JSON exactly in this schema: {"updated_memories":{"
    memory_id":{"updated_text":"..."}}}.
```

B.2.4 HIERARCHICAL DECISIONS AND INTER-AGENT HANDOFF

Shared Decision Prompt

```
YOU ARE {AGENT_NAME}.
Decide whether the updated current-level dynamic memory is sufficient
     or whether to descend. Do not answer the question.
{LEVEL_SPECIFIC_INSTRUCTION}
Return strict JSON with keys: selected_child_memory_ids,
    stop_descending, next_instruct, reason, template_type. If
    stop_descending is true, selected_child_memory_ids must be [] and
     next_instruct must be null. If continuing,
    selected_child_memory_ids must be current-level memory ids
    supplied in updated_dynamic_memories.
```

Hierarchical Decision Instructions

If $\ell = L_0$ (Super), decide $L_0 \to L_1$:

```
If continuing to Macro, generate next_instruct using one of these L0
    ->L1 types: main_idea, temporal_order, identity,
    action_verification, spatial_relation, causal_result. The
    next_instruct must be one sentence that tells Skim Agent what
    type of Macro information to find. Return template_type as the
    selected type. Only global main-idea/title/summary questions may
    stop at Super. All other question types must continue to Macro.
```

Else if $\ell = L_1$ (Macro), decide $L_1 \to L_2$:

```
If continuing to Subgraph, generate a short instruction in the form:
    Found: [...]; still need to verify: [...]. For non-global detail
    questions, prefer continuing to Subgraph after at least one Macro
     observation; old Macro text alone is not enough to terminate.
```

Else if $\ell = L_2$ (Subgraph), return STOP:

```
Focus is the final retrieval layer; stop_descending must be true.
```

B.2.5 FINAL ANSWER AGENT

Final Answer Agent System Prompt

```
YOU ARE THE FINAL ANSWER AGENT.
Based on the memory provided below, return strict JSON with keys:
    reasoning, answer.
```

Final Answer Agent User Prompt

```
Question:
{RAW_QUESTION_WITH_OPTIONS}

Retrieved video memory:
{SERIALIZED_QUERY_ADAPTIVE_MEMORY}
```

C Memory Ablation Protocol

Shared evaluation setting. All configurations in Table 3(b) use the same 200-question subset of Video-MME Long, GPT-5.5, and the same preconstructed hierarchical memory. We distinguish the static-retrieval baseline from the controlled comparison of final memory components.

Static retrieval baseline. *Static Memory Only* performs hierarchical retrieval without Visual, ASR, or Update calls. The Answer Agent receives the retrieved static SuperEvent, MacroEvent, and Subgraph contents together with the retrieval trace from that run; node contents are neither compressed nor rewritten. Its comparison with VideoTapestry therefore measures the combined effect of query-time observation and memory refinement rather than memory composition alone.

Controlled final-memory ablations. For each question, VideoTapestry first completes its full query-time exploration. The resulting selected nodes, observations, refined contents, and exploration trace are then fixed, and three final memory representations are constructed:

- *Refined Super Only* retains the refined Super-level scaffold while excluding MacroEvent and Subgraph contents.
- *Refined Paths Only* retains refined MacroEvent and Subgraph contents along selected paths while excluding SuperEvents.
- *VideoTapestry* combines the refined Super-level scaffold with the refined paths.

The three variants use the same exploration trace and do not rerun retrieval, observation, or updating; only the memory contents supplied to the Answer Agent differ. Thus, their comparison isolates the contribution of the final serialized memory components under a fixed exploration result.

D Limitations and Future Work

VideoTapestry is evaluated with a fixed, preconstructed video hierarchy. Although query-time observations can supplement missing details, retrieval remains constrained by the initial segmentation and parent–child structure. Incorrect event grouping or temporal boundaries may therefore hinder access to relevant evidence. Evaluating the framework with alternative memory constructors and allowing selective structural revision are promising directions.

The framework also incurs offline memory-construction costs and additional query-time model and tool calls. The reported accuracy gains therefore do not establish a computational-efficiency advantage. Future work could investigate budget-aware observation and refinement while separately accounting for offline and per-query costs.

Finally, query-time refinement relies on model-generated interpretations of visual and transcript evidence, whose errors may propagate into refined memory and subsequent decisions. Our prompts encourage uncertainty reporting and evidence-supported updates, while Super-level memory records visual and transcript-derived observations separately. These safeguards do not guarantee correct interpretation or fully resolve conflicts between stored and newly observed information. Future work could track evidence provenance across levels and selectively re-verify uncertain or conflicting updates.