VIDEOPURSUIT: FROM RETRIEVAL ACTIONS TO LOCAL RETRIEVAL PROCESSES FOR AGENTIC LONG-VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Agentic long-video understanding increasingly relies on adaptive video retrieval to access sparse, task-relevant observations during reasoning. However, many existing methods follow an action-level handoff pattern, returning control to global reasoning after each retrieval action. A useful retrieval action may still only narrow a temporal region, reveal a new cue, or indicate what should be inspected next, so completing the action does not necessarily mean that the current retrieval is ready for handoff. We introduce VideoPursuit, a framework that treats retrieval for each local objective as a dedicated process rather than a sequence of retrieval actions individually managed by global reasoning. When global reasoning over the original question identifies an information gap, it delegates a local retrieval objective to a dedicated retrieval agent. Within a bounded local execution scope, the agent may perform multiple retrieval actions around the same objective, using explicit local audit after each action to determine whether further retrieval is needed or the current result is ready for handoff. Intermediate retrieval states remain local, while acquired observations and their source information are selectively returned to global reasoning. Across four long-video understanding benchmarks, VideoPursuit achieves an average accuracy of 82.0%, improving by 3.4 percentage points over the strongest prior agentic system. Ablations further show that persistent local retrieval substantially outperforms action-level return, while selective information propagation across the local–global boundary provides additional gains.

1 INTRODUCTION

Long-video understanding requires reasoning over events and relationships whose relevant information may be sparsely distributed across extended temporal spans. Exhaustively processing an entire video is expensive, while compressed or sparsely sampled representations may miss critical details (Song et al., 2024; Shen et al., 2025; Wu et al., 2024). Inspired by advances in adaptive information acquisition in broader agent systems (Li et al., 2025; Wan et al., 2026b; Wang et al., 2026a; Li et al., 2026b), recent works have explored iterative video retrieval to locate and access task-relevant visual segments or transcript content during long-video understanding (Wang et al., 2025b; Zhang et al., 2025; Yuan et al., 2026). Some methods further incorporate reflection or explicit verification to improve the quality of acquired information (Wang et al., 2026c;b; Yan et al., 2026a). By selectively accessing fine-grained or temporally localized information that may be absent from the initially available context, these approaches have improved long-video understanding without requiring exhaustive processing of the entire video.

However, in these systems, the retrieval loop is primarily controlled by global reasoning, which works on the original question. Recent methods do revisit their retrieval: a reflection or verification step judges whether the evidence so far is sufficient and plans another round if it is not (Wang et al., 2026c; Yan et al., 2026a; Wang et al., 2026b). The process, however, focuses primarily on the original question rather than the gap that prompted the retrieval, and the intermediate products of retrieval flow back into the context driving later rounds. This matters in long-video understanding, where a single retrieval action exposes only a limited portion of an extended temporal input. An action that narrows a temporal range or surfaces a new cue is useful, yet leaves the local objective unsettled, and a question-level criterion cannot tell how close it is to being resolved. This raises two

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.

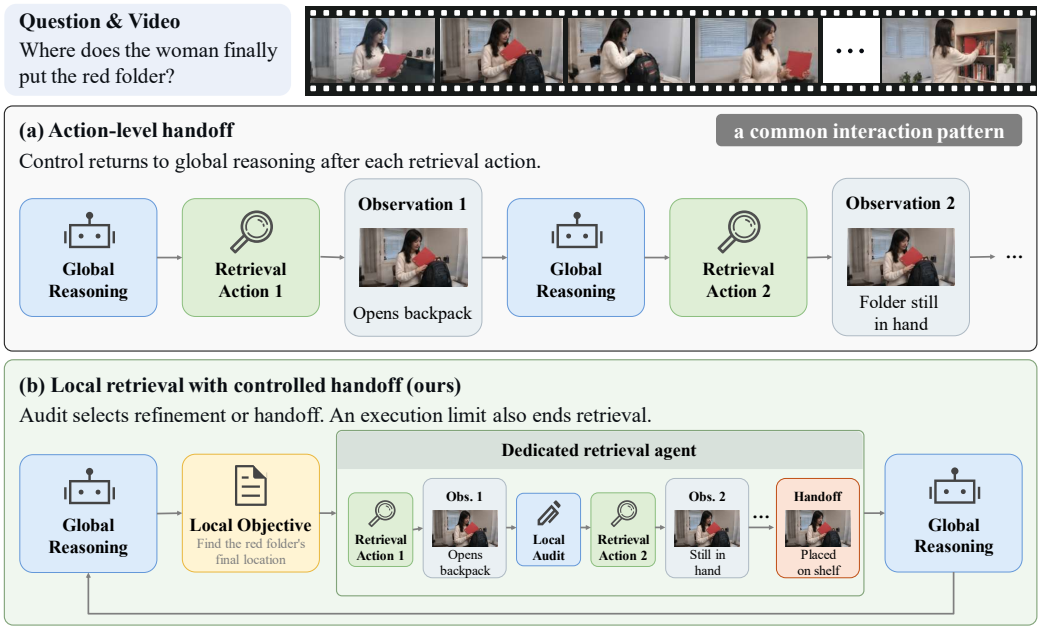

Figure 1: Illustration of action-level handoff and controlled local retrieval. (a) Global reasoning resumes after each retrieval action. (b) A dedicated agent refines retrieval within a local scope and returns observations with compact terminal information when the local process ends.

challenges. First, the system must judge when the current retrieval is adequate for handoff, and that judgment has to be made with respect to the local objective rather than solely the original question. Such refinement must also remain bounded rather than continue indefinitely. Second, when further retrieval is needed, it must be carried out without repeatedly returning control to global reasoning. Otherwise global reasoning has to advance the original question and manage an unfinished local objective at the same time. Tentative interpretations, unsuccessful attempts, and follow-up plans then enter later task-level decisions, where local errors can be amplified over a long reasoning trajectory.

We address these challenges with VideoPursuit, a framework that organizes each local retrieval objective as a pursuit: a dedicated local retrieval process with explicit local audit and controlled handoff (Figure 1). When global reasoning identifies an information gap, it formulates a local retrieval objective and delegates it to a dedicated retrieval agent. The agent performs retrieval actions around this objective and applies a local audit after each action to decide whether the result is adequate for handoff. If it is not, the agent continues within the local scope; otherwise, or when the bounded execution limit is reached, control returns. Intermediate retrieval states remain local, and only the acquired observations together with compact terminal information are returned, so reasoning over the original question continues without inheriting them.

We evaluate our approach on four long-video understanding benchmarks: LVBench, MLVU, Video-MME Long, and LongVideoBench (Wang et al., 2025a; Zhou et al., 2025; Fu et al., 2025; Wu et al., 2024). VideoPursuit reaches an average accuracy of 82.0%, 10.4 percentage points above direct inference with the same backbone and 3.4 above AVP (Wang et al., 2026c), the strongest prior agentic system reported on all four benchmarks. Ablations show that persistent local retrieval is not recovered by returning local status to a question-level loop, and that the local–global information boundary contributes independently.

Our contributions are threefold:

1. We identify a control gap in agentic long-video understanding: because retrieval is judged against the original question, neither completing an action nor finding the accumulated evidence insufficient indicates whether the local objective has been resolved.

2. We introduce VideoPursuit, which delegates each local objective to a dedicated agent that spans multiple retrieval actions within a bounded scope, judges adequacy for handoff locally, and returns only observations and compact terminal information.
3. We evaluate the framework on four long-video understanding benchmarks, reaching 82.0% average accuracy, 10.4 percentage points above the same backbone and 3.4 above the strongest prior agentic system, with ablations isolating both mechanisms.

2 Related Work

2.1 Agentic Long-Video Understanding

Agentic approaches to long-video understanding allow models to actively retrieve task-relevant observations during reasoning rather than rely solely on a preconstructed video representation. Early methods enable selective access to long videos through visual tools, structured memory, and retrieval mechanisms (Wang et al., 2025b; Fan et al., 2025; Zhang et al., 2024; Shang et al., 2024). Subsequent work further develops reasoning-time video interaction. Deep Video Discovery invokes video tools according to the current reasoning state (Zhang et al., 2025), while VideoExplorer organizes exploration through hierarchical orchestration and task-oriented perception (Yuan et al., 2026). LensWalk adaptively controls temporal scope and sampling density (Li et al., 2026a), and LongVideoAgent coordinates specialized components for long-video reasoning (Liu et al., 2026b).

More recent methods increasingly incorporate reflection and verification into this iterative process. VideoHV-Agent structures long-video reasoning around hypothesis verification (Wang et al., 2026b), while AVP uses a plan–observe–reflect loop to assess whether the currently acquired observations are sufficient for answering the question (Wang et al., 2026c). Reflect-R1 explores verification-driven self-correction (Chen et al., 2026), and REVEAL diagnoses insufficient information and triggers targeted re-retrieval (Yan et al., 2026a). VideoSEAL further separates planning from answer authority during extended video exploration (Qiu et al., 2026). Together, these works establish that long-video reasoning can benefit from both repeated retrieval and explicit reassessment of acquired information.

Our focus is complementary. In many existing systems, an individual retrieval, localization, or inspection action returns its observation to question-level reasoning, which then determines whether and how retrieval should continue (Wang et al., 2025b; Zhang et al., 2025; Wang et al., 2026c; Yan et al., 2026a). A useful observation, however, may still leave the current retrieval suboptimal by only narrowing a temporal region, revising an interpretation, or revealing what should be inspected next. We therefore study how multiple retrieval actions can be organized into a local retrieval process around the same objective, and when that process is adequate for handoff to global reasoning.

2.2 Action Boundaries in Long-Horizon Agents

Related questions have also emerged in long-horizon agent research. Deep-research systems commonly interleave reasoning with repeated retrieval, navigation, verification, and refinement rather than treating retrieval as a single call. WebThinker dynamically gathers external information when reasoning exposes knowledge gaps (Li et al., 2025), while DeepVerifier uses test-time verification feedback to iteratively refine the research process (Wan et al., 2026b). Recent video deep-research work extends similar ideas to video-conditioned open-world reasoning by combining video grounding with external retrieval and verification (Liu et al., 2026a;c; Zhang et al., 2026). These works reflect a broader recognition that information acquisition in long-horizon tasks may require sustained interaction between reasoning and retrieval rather than isolated retrieval calls.

From the perspective of action execution in agent systems, Action Boundary Blindness examines failures in identifying appropriate action granularity, scope, and completeness (Wang et al., 2026a). Voyager organizes reusable behaviors as skills (Wang et al., 2024), while ReAcTree allows local execution to persist across multiple reasoning and acting steps under explicit control flow (Choi et al., 2026). These studies similarly highlight that the appropriate unit of execution may extend beyond a single primitive interaction.

Long-horizon agent research has also increasingly considered how intermediate execution states should participate in subsequent reasoning. HiAgent, AgentFold, COMPASS, and Sculptor orga-

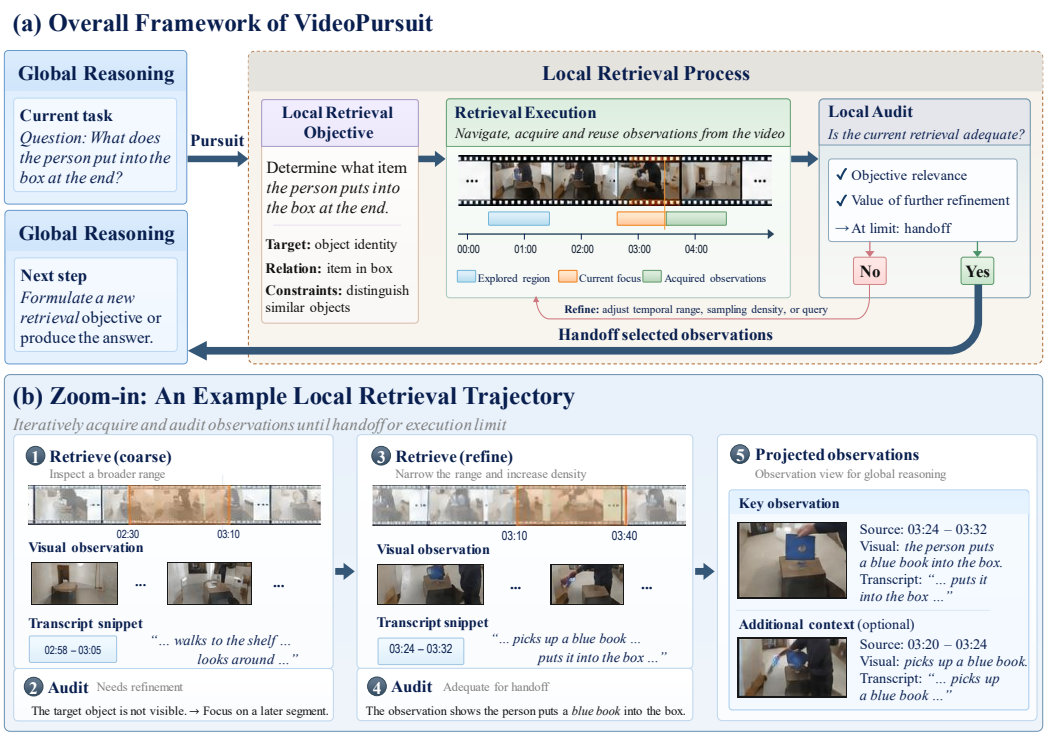

Figure 2: Overview of VideoPursuit. (a) Global reasoning and local retrieval. (b) An example local retrieval trajectory, illustrating how VideoPursuit progressively gathers, audits, and refines observations before handoff.

nize, consolidate, or selectively expose intermediate execution states rather than retaining the entire interaction history (Hu et al., 2025; Ye et al., 2026; Wan et al., 2026a; Li et al., 2026b). These studies reflect a broader concern with controlling how intermediate states are propagated across extended reasoning trajectories.

These lines of work provide broader precedents for structuring multi-step execution and its boundaries in long-horizon agents. Our focus is the specific interface between video retrieval and global reasoning in agentic long-video understanding: how multiple retrieval actions around the same local objective should be organized before handoff, when the resulting retrieval is adequate to return, and which retrieval-specific states should participate in subsequent global reasoning.

3 METHODOLOGY

Given a long video V, a question q, and answer candidates C, VideoPursuit (Figure 2) separates global reasoning over the original task from the local retrieval process used to acquire missing information. A Controller maintains the global reasoning state and identifies information gaps that remain relevant to the original task. For each such gap, global reasoning formulates a local retrieval objective and initiates a pursuit, i.e., a dedicated local retrieval process executed by a retrieval agent within a local execution scope. Within this scope, the retrieval agent acquires observations, maintains a local working state, and determines whether the current retrieval is adequate for handoff or would benefit from further refinement. Control returns to global reasoning only after the local retrieval process terminates, rather than automatically after each retrieval action.

We denote the global reasoning trajectory by

$$\tau^G = \left(G_0 \xrightarrow{\tau_0^L} G_1 \xrightarrow{\tau_1^L} \cdots \xrightarrow{\tau_{T-1}^L} G_T \right), \tag{1}$$

where G_t is the t-th global reasoning state, τ_t^L is the local retrieval trajectory initiated at G_t, and T is the number of completed local retrieval processes. Each transition represents the local execution between two successive global reasoning states; it does not imply that the complete local trajectory is inserted into the global reasoning context. The local process receives its retrieval objective and relevant observations rather than the complete question-level reasoning trace. Conversely, when local execution ends, the Controller receives the acquired observations rather than the local working trajectory. This separation allows the acquired observations to inform subsequent global reasoning while keeping intermediate retrieval interpretations and plans within the local execution scope.

3.1 Global Reasoning and Local Retrieval Objectives

Global reasoning considers the original question and answer candidates using the observations currently available to it. It maintains task-level assessments that may be revised as additional observations become available. After a local retrieval process hands control back, global reasoning reassesses the original task and determines whether further information is required.

When global reasoning identifies an information gap, it formulates a local retrieval objective r_t, expressed as a natural-language proposition specifying a fact or relation to be established through video retrieval. The objective specifies what information should be established, rather than prescribing where or how to obtain it. In particular, it does not specify a fixed temporal interval, retrieval sequence, observation source, or retrieval configuration. These decisions are delegated to the retrieval agent within its local execution scope.

Global reasoning therefore retains responsibility for the original question, while the retrieval agent operates around the local objective within a dedicated local execution scope. Importantly, local retrieval adequacy and global answer sufficiency are distinct decisions. The local audit determines whether the current retrieval is adequate for handoff, whereas global reasoning independently determines whether the accumulated observations provide sufficient support for answering the original question. The pursuit may therefore terminate even when the original task still requires additional information. In that case, global reasoning can identify another information gap, formulate a new local retrieval objective, and initiate another pursuit.

3.2 Local Retrieval Process

For each local retrieval objective r_t, the retrieval agent operates within a dedicated local execution scope that persists across retrieval actions for the same objective. Its decisions are conditioned on the active objective, available observations, retrieval capabilities, and a local working state, rather than the complete question-and-option reasoning context. The local working state records information needed to continue the current retrieval process, including previous attempts, tentative interpretations, remaining uncertainties, and intermediate retrieval plans.

We represent the corresponding local retrieval trajectory as

$$\tau_t^L = \left(L_{t,0} \xrightarrow{u_{t,1}} O_{t,1} \xrightarrow{a_{t,1}} L_{t,1} \to \cdots \to L_{t,J_t-1} \xrightarrow{u_{t,J_t}} O_{t,J_t} \xrightarrow{a_{t,J_t}} \mathrm{END} \right), \tag{2}$$

where $L_{t,0}$ is the initial local working state and $L_{t,j}$ is the state retained after a nonterminal local step. $u_{t,j}$ denotes a retrieval action, $O_{t,j}$ denotes the observation material made available by that action, and $a_{t,j}$ denotes the subsequent local audit decision. J_t is the terminal local step. Observation material may include newly acquired or previously available content, so a local step does not necessarily create a new observation record.

Retrieval Execution. We construct a query-independent temporal index over contiguous video segments represented with OpenCLIP features for navigation (Radford et al., 2021; Ilharco et al., 2021). The active local retrieval objective retrieves candidate temporal regions as navigation cues without restricting the retrieval agent to those regions. At each local step, the agent selects an observation source, an immediate retrieval target, and the corresponding retrieval configuration according to the active objective and its current working state. Our implementation supports visual observation, transcript access, and reuse of previously acquired observations. The local retrieval objective remains fixed throughout the process, while the immediate retrieval target may be revised after each observation. A newly inspected segment may narrow the relevant temporal region, reveal another source to

inspect, or prompt revision of a tentative interpretation. Such information guides the next retrieval action within the same pursuit without requiring an intervening global-reasoning step. Acquired observations retain both their content and source information, allowing subsequent reasoning to trace them to the inspected video or transcript material. Relevant observations obtained during earlier retrieval processes remain reusable in later retrieval and reasoning. The execution scope therefore constrains the retrieval-specific working process rather than the reuse of observations: later local processes may use previously acquired information without inheriting the complete working trajectory that produced it.

Local Audit. After each retrieval action, the local audit evaluates the current retrieval state with respect to the active objective, acquired observations, previous local attempts, and available retrieval capabilities. It assesses whether the acquired observations support a response to the active objective, identifying unsupported links and unresolved facts. If further refinement is beneficial, control remains within the local retrieval process. Audit feedback updates the local working state and guides the next retrieval action, which may revise the observation source, temporal range, or immediate retrieval target. The feedback specifies the unresolved fact or relation, while the retrieval agent determines which observation source and acquisition parameters can address that gap. Global reasoning is not resumed between these local steps. When the audit judges the current retrieval adequate for handoff, the local process terminates normally. The audit may also end retrieval when available capabilities cannot resolve the remaining gap. The process may also terminate when its bounded execution limit is reached. These outcomes are model-mediated control judgments about whether to return control, rather than independent measurements of retrieval quality or guarantees of observation correctness. A compact terminal status can optionally indicate whether the process ended through normal handoff or an execution limit, but this status is not treated as a separate reasoning object. Thus, completing an individual retrieval action does not itself determine when control returns; the local audit determines whether retrieval should continue or whether its current result should be handed back to global reasoning.

3.3 Controlled Handoff and Observation Projection

A local retrieval trajectory produces more than observations. It also generates tentative interpretations, retrieval plans, unsuccessful attempts, audit feedback, and other working states that help the retrieval agent pursue its local objective. These records are useful for local execution but are not automatically transferred as premises for reasoning about the original question.

When the local process terminates, the system constructs an observation view from the accumulated retrieval records. Let M_{t+1} denote the cumulative records available after the t-th local process terminates, including acquired observations and local working records. The next global reasoning state is obtained as

$$G_{t+1} = U_G\left(G_t,\, \Pi_{\text{obs}}(M_{t+1})\right), \tag{3}$$

where Π_{obs} constructs the observation view exposed to the Controller and U_G denotes the subsequent global reasoning update.

The projection retains observation content together with its source information and relevant qualifications, while excluding local retrieval plans, audit rationales, tentative local conclusions, and retrieval-process history. These local records may remain stored for execution or inspection, but their persistence does not make them part of the Controller's reasoning input. Conversely, acquired observations remain reusable across local retrieval processes, and the Controller may retain and revise its own task-level assessments.

The separation therefore concerns which information participates in which reasoning process, rather than requiring physically isolated storage or discarding previously acquired observations. Using the updated observation view, the Controller reassesses the original question and answer candidates. If another information gap remains relevant, it formulates a new local retrieval objective and initiates another local process; when sufficient task-level support is available, it proceeds to answer.

Local retrieval adequacy and global answer sufficiency therefore remain explicitly separated: the former determines when one local retrieval process hands control back, whereas the latter determines when reasoning over the original task can terminate.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmarks and Evaluation. We evaluate our method on four representative long-video question answering benchmarks: LVBench (Wang et al., 2025a), LongVideoBench (Wu et al., 2024), Video-MME Long (Fu et al., 2025), and MLVU (Zhou et al., 2025). Following the multiple-choice protocol commonly adopted in recent agentic long-video studies (Wang et al., 2026c; Yan et al., 2026b), we report accuracy on all benchmarks. Specifically, we evaluate on the standard LVBench setting, the public validation split of LongVideoBench, the Long split of Video-MME, and the multiple-choice test samples of MLVU. All main experiments use speech information obtained through offline ASR preprocessing with WhisperX (Bain et al., 2023).

Baselines. We compare against general-purpose multimodal models and representative long-video and agentic video understanding systems, including methods based on active perception, retrieval and memory, tool-augmented reasoning, and multi-agent coordination (Wang et al., 2026c; Zhang et al., 2025; Yan et al., 2026b;a; Qiu et al., 2026). Table 1 presents the comparison results alongside the backbone or core model used by each method. We additionally evaluate Qwen3.8-27B under our evaluation pipeline using 256 uniformly sampled frames and the same WhisperX ASR input.

Implementation Details. We use Qwen3.8-27B (Qwen Team, 2026) as the reasoning backbone and keep the same model and inference configuration across all four benchmarks. Each video is initialized with a coarse overview of 128 uniformly sampled frames together with its offline WhisperX transcript. For subsequent retrieval, we construct an OpenCLIP-based (Ilharco et al., 2021) temporal index over the video. Given a local retrieval objective, the index provides candidate temporal regions as navigation cues, while the retrieval process remains free to inspect other valid intervals or broaden its observation when necessary. The process can inspect selected video regions or the full video, access the transcript, and reuse previously acquired observations. Additional implementation details are provided in the appendix.

4.2 MAIN RESULTS

Table 1 compares our method with representative direct multimodal models and recent agentic long-video systems. Using the same Qwen3.8-27B backbone and configuration across all four benchmarks, our method achieves 81.5% on LVBench, 76.3% on LongVideoBench, 85.8% on Video-MME Long, and 84.4% on MLVU, corresponding to an average accuracy of 82.0%. Under our evaluation pipeline, VideoPursuit improves average accuracy by 10.4 percentage points compared with direct Qwen3.8-27B using the same reasoning backbone, while using fewer frames on average than the uniform baseline.

Compared with AVP (Wang et al., 2026c), a video-input-based agentic method with the highest four-benchmark average accuracy among the baselines in Table 1, VideoPursuit improves the average by 3.4 percentage points. Among methods operating under the same multi-frame video-input setting as VideoPursuit, Symphony (Yan et al., 2026b) is the strongest baseline in Table 1, and VideoPursuit improves over it by 5.0 percentage points. The gains are particularly pronounced on LVBench and Video-MME Long: our method reaches 81.5% on LVBench, compared with 76.0% for DVD with auxiliary transcript input (Zhang et al., 2025), and achieves 85.8% on Video-MME Long, compared with 81.9% for AVP (Wang et al., 2026c). These results show that the proposed framework maintains strong performance across heterogeneous long-video QA benchmarks with a unified backbone and inference configuration.

4.3 ABLATION STUDIES

We examine the two design choices central to our framework: persistent local retrieval and the information boundary between local retrieval and global reasoning. Unless otherwise specified, all variants use the same model and implementation as the full method and modify only the corresponding control-flow or information-flow design.

Table 1: Comparison with existing methods on four long-video QA benchmarks.

Method	Reasoning Model	LVBench	LongVideo Bench	Video-MME Long	MLVU
Direct MLLMs					
Seed 1.6 VL (ByteDance Seed Team, 2025)	Seed 1.6 VL	58.1	66.1	68.4	65.3
Gemini-2.5-Pro (Comanici et al., 2025)	Gemini-2.5-Pro	78.7	–	84.3	–
Qwen3-VL (Bai et al., 2025)	Qwen3-VL	67.7	–	–	84.3
Qwen3.8-27B† (Qwen Team, 2026)	Qwen3.8-27B	67.2	72.4	70.3	76.7
Agentic / Long-Video Methods					
VideoLucy (Zuo et al., 2025)	DeepSeek-R1	58.8	–	66.8	76.1
Video-RAG-72B (Luo et al., 2025)	LLaVA-Video-72B	–	65.4	73.1	73.8
LVAgent (Chen et al., 2025)	Multi-agent MLLMs	–	80.0	81.7	83.9
DVD (Zhang et al., 2025)	OpenAI o3	74.2	71.6	67.3	–
DVD + Auxiliary Transcript (Zhang et al., 2025)	OpenAI o3	76.0	–	–	–
VideoExplorer (Yuan et al., 2026)	Qwen2.5-7B	51.4	–	–	58.6
GOPAgen (Chi et al., 2026)	DeepSeek-V3	73.5	73.2	66.7	77.3
REVEAL-27B (Yan et al., 2026a)	Qwen3.5-27B	65.9	75.7	79.1	–
VideoSEAL + Kimi2.5 (Qiu et al., 2026)	Kimi2.5	65.7	80.8	–	84.3
Symphony (Yan et al., 2026b)	DeepSeek-R1/V3	71.8	77.1	78.1	81.0
AVP (Wang et al., 2026c)	Gemini-2.5-Pro	74.8	73.4	81.9	84.3
VideoPursuit	Qwen3.8-27B	81.5	76.3	85.8	84.4

Baseline scores are taken from published results; Seed 1.6 VL scores follow Symphony (Yan et al., 2026b).
† Evaluated by us with 256 uniformly sampled frames and WhisperX ASR.

4.3.1 Local Retrieval Control

We compare three retrieval-control strategies (Figure 3(a)). Action Return hands control back to global reasoning after every retrieval action. Terminate Status additionally reports whether further local refinement would be beneficial, while still returning after every action. Persistent Local Retrieval, our full method, instead uses the same local assessment to control execution within the retrieval process, allowing retrieval to continue around the same objective before handoff.

Action Return achieves 64.0%, while adding local-status feedback improves accuracy to 68.7%. Persistent local retrieval further increases accuracy to 81.1%. Notably, all variants share the same question-level resource ceiling, with comparable average numbers of acquisition actions across variants. The large remaining gap shows that explicitly communicating local status cannot substitute for keeping the retrieval process local and allowing subsequent retrieval actions to build directly on newly acquired observations.

4.3.2 Information Boundaries

We next examine information propagation across the local–global boundary while preserving the same persistent local retrieval process (Figure 3(b)). Global Context additionally exposes the complete global-reasoning state to local retrieval, whereas Local Trace exposes the complete local working trajectory to subsequent global reasoning after handoff.

Both variants underperform the full method: exposing the complete global context to local retrieval achieves 77.1%, while returning the complete local trajectory to global reasoning achieves 75.7%, compared with 81.1% for the selective interface used by our full method. These results indicate that unrestricted information sharing in either direction is unnecessary; restricting each reasoning process to the information relevant to its role provides a more effective local–global interface.

	Setting	Acc. (%)	Δ (pp)
Full	**VideoPursuit**	**81.1**	0.0
(a) Local retrieval control			
A1	Action Return	64.0	−17.1
A2	Terminate Status	68.7	−12.4
(b) Information boundary			
A3	Global Context	77.1	−4.0
A4	Local Trace	75.7	−5.4

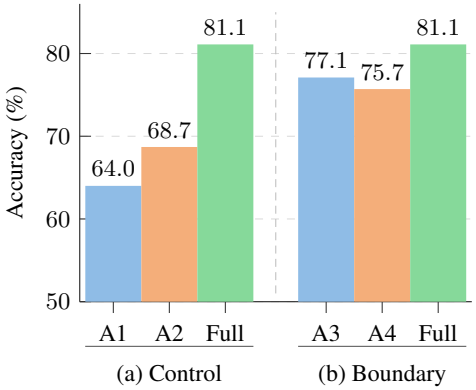

Figure 3: Ablations of VideoPursuit. **Left:** Accuracy and differences from Full (Δ, in percentage points). **Right:** (a) Local retrieval control. (b) Information boundary. Full denotes VideoPursuit.

4.4 Qualitative Analysis

We examine two cases to illustrate the roles of persistent local retrieval and selective information propagation. Detailed trajectories and source information are provided in Appendix C.

Persistent local retrieval. The LVBench case asks who turns first during the third contestant's performance. Answering requires establishing which performance is third before comparing the chair turns within it. A correct comparison of chair turns within a segment is therefore insufficient if that segment has not been linked to the requested performance. VideoPursuit maintains the same objective throughout the local retrieval process while resolving conflicting candidate reports, establishing performance order, and inspecting the relevant chair turns. The local objective provides a consistent reference for deciding what should be inspected next. The returned observations identify the person in pink as the first to turn, supporting the correct answer. The Action-Level Return run instead inspects a later performance while retaining an earlier clothing interpretation and ultimately selects the person in a black jacket. This example illustrates how an unresolved temporal relation can guide successive local acquisitions and how intermediate observations can inform the next inspection before control returns to global reasoning.

Selective information propagation. The MLVU case asks how many cookie-baking scenes appear. Both runs acquire observations covering ingredient preparation and a later oven interaction, but their local audits interpret baking narrowly and exclude preparation. The observations describe the activities present in the video, while the count additionally depends on which of those activities are included in the requested category. In VideoPursuit, the observations are returned without the local count of one. Global reasoning then reconsiders the requested category and counts both reported segments, yielding the reference answer of two. The Full Local Trace → Global run also receives the local interpretation and retains the count of one. This case illustrates why observations can remain useful even when a local interpretation is too narrow: selective propagation preserves the opportunity to reconsider their meaning with respect to the original question.

5 Conclusion

We presented VideoPursuit, a local retrieval framework for agentic long-video understanding that allows retrieval to continue around a local objective before handing control back to global reasoning. The framework treats retrieval for each local objective as a pursuit, a dedicated process that uses local audit to assess adequacy for handoff, keeps retrieval-specific working states local within its execution scope, and selectively exposes acquired observations to subsequent reasoning. Across four long-video QA benchmarks, our approach achieves an average accuracy of 82.0%, improving by 3.4 percentage points over the leading agentic method. Ablations further support both persistent local retrieval and selective information propagation across the local–global boundary. These results suggest that effective agentic video reasoning depends not only on what information is retrieved, but also on how the retrieval process is organized before its results re-enter global reasoning.

AI USE STATEMENT

During this work, we used generative AI tools to assist with literature discovery and summarization, code implementation and debugging, analysis of experimental outputs, figure preparation, and drafting and revising selected manuscript passages. Generative AI was also used as a discussion aid for conceptual framing and experimental planning. The authors formulated the research problem and core methodological ideas, designed the experimental and ablation studies, made the research and implementation decisions, interpreted the experimental findings, and substantially revised and finalized the manuscript. AI-assisted implementation was checked through targeted tests, and numerical results were cross-checked against recorded experimental outputs. The authors reviewed all AI-assisted content and take full responsibility for the final work.

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.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. In *Interspeech 2023*, pp. 4489–4493, 2023. doi: 10.21437/Interspeech.2023-78.

ByteDance Seed Team. Seed1.6: Tech introduction. Official technical introduction, June 2025. URL https://seed.bytedance.com/en/seed1_6. Published June 25, 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, 2025.

Shuimu Chen, Yuteng Chen, Yuanshen Guan, Zebang Cheng, Zeyu Zhang, Shengqian Qin, Bin Xia, Jiaran Li, Wenming Yang, and Fei Ma. Reflect-R1: Evidence-driven reflection for self-correction in long video understanding. In *Computer Vision – ECCV 2026*, 2026. URL https://arxiv.org/abs/2606.27922.

Haozhe Chi, Yang Jin, and Yadong Mu. Gopagen: Motion-aware and efficient agentic long-video understanding with structural memory and hierarchical reasoning, 2026. URL https://arxiv.org/abs/2606.06532.

Jae-Woo Choi, Hyungmin Kim, Hyobin Ong, Youngwoo Yoon, Minsu Jang, Dohyung Kim, and Jaehong Kim. Reactree: Hierarchical llm agent trees with control flow for long-horizon task planning, 2026. URL https://arxiv.org/abs/2511.02424.

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.

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 Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 75–92, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72670-5.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, June 2025.

Mengkang Hu, Tianxing Chen, Qiguang Chen, Yao Mu, Wenqi Shao, and Ping Luo. HiAgent: Hierarchical working memory management for solving long-horizon agent tasks with large language model. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar

(eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32779–32798, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1575. URL https://aclanthology.org/2025.acl-long.1575/.

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773.

Keliang Li, Yansong Li, Hongze Shen, Mengdi Liu, Hong Chang, and Shiguang Shan. Lenswalk: Agentic video understanding by planning how you see in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19518–19528, June 2026a.

Mo Li, L.H. Xu, Qitai Tan, Long Ma, Hongyong Song, Ting Cao, and Yunxin Liu. Sculptor: Empowering llms with cognitive agency via active context management. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 153411–153440, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/f9155a06ba1b20e40c38bf8541db456f-Paper-Conference.pdf.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. Webthinker: Empowering large reasoning models with deep research capability. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 120091–120131. Curran Associates, Inc., 2025. doi: 10.52202/085713-4011. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/ae03bdef276132fae089692445725635-Paper-Conference.pdf.

Chengwen Liu, Xiaomin Yu, Zhuoyue Chang, Zhe Huang, Shuo Zhang, Heng Lian, Jisheng Dang, Rui Xu, Sen Hu, Jianheng Hou, Chengwei Qin, Xiaobin Hu, Kunyi Wang, Zhi Yang, Hao Peng, Hong Peng, Ronghao Chen, and Huacan Wang. Watching, reasoning, and searching: A video deep research benchmark on open web for agentic video reasoning, 2026a. URL https://arxiv.org/abs/2601.06943.

Runtao Liu, Ziyi Liu, Jiaqi Tang, Yue Ma, Renjie Pi, Jipeng Zhang, and Qifeng Chen. LongVideoAgent: Multi-agent reasoning with long videos. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40404–40416, San Diego, California, United States, July 2026b. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1876. URL https://aclanthology.org/2026.acl-long.1876/.

Wenqi Liu, Shijie Ma, Yunxiao Wang, Meng Liu, Qile Su, Han Liu, Bohan Hou, Zeyu Wang, Xuanyu Zheng, Changyi Liu, Tianke Zhang, Haonan Fan, Kaiyu Jiang, Yingxin Li, Jiankang Chen, Xu Wang, Hongyi Fu, Jianxiong Wang, Bin Wen, Tingting Gao, Han Li, Jianhua Yin, Yinwei Wei, and Xuemeng Song. Thinking beyond videos: Unifying video reasoning and deep research for open-world video agents, 2026c. URL https://arxiv.org/abs/2608.23329.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. Video-RAG: Visually-aligned retrieval-augmented long video comprehension. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=QaZxGWlbgO.

Chenhao Qiu, Yechao Zhang, Xin Luo, Shien Song, and Xusheng Liu. VideoSEAL: Mitigating evidence misalignment in agentic long video understanding by decoupling answer authority. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=RsPequeaiZ.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Chuyi Shang, Amos You, Sanjay Subramanian, Trevor Darrell, and Roei Herzig. TraveLER: A modular multi-LMM agent framework for video question-answering. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9740–9766, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.544. URL https://aclanthology.org/2024.emnlp-main.544/.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal adaptive compression for long video-language understanding. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54582–54599. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/shen25j.html.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18221–18232, June 2024.

Guangya Wan, Mingyang Ling, Xiaoqi Ren, Rujun Han, Sheng Li, and Zizhao Zhang. COMPASS: Enhancing agent long-horizon reasoning with evolving context. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3360–3380, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.152. URL https://aclanthology.org/2026.acl-long.152/.

Yuxuan Wan, Tianqing Fang, Zaitang LI, Yintong Huo, Wenxuan Wang, Haitao Mi, Dong Yu, and Michael R. Lyu. Inference-time scaling of verification: Self-evolving deep research agents via test-time rubric-guided verification. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 24822–24835, San Diego, California, United States, July 2026b. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1243. URL https://aclanthology.org/2026.findings-acl.1243/.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An Open-Ended Embodied Agent with Large Language Models. *Transactions on Machine Learning Research*, 2024. URL https://mlanthology.org/tmlr/2024/wang2024tmlr-voyager/.

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

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 58–76, Cham, 2025b. Springer Nature Switzerland. ISBN 978-3-031-72989-8.

Zhangyi Wang, Bingnan Yu, Jiexiang Xu, and Zongze Li. Action boundary blindness: When LLM agents cannot tell where one action ends and another begins. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36883–36899, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1711. URL https://aclanthology.org/2026.acl-long.1711/.

Zheng Wang, Haoran Chen, Haoxuan Qin, Zhipeng Wei, Tianwen Qian, and Cong Bai. Think, then verify: A hypothesis-verification multi-agent framework for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33784–33793, June 2026b.

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 (CVPR) Findings*, pp. 9088–9099, June 2026c.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 28828–28857. Curran Associates, Inc., 2024. doi: 10.52202/079017-0907. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/329ad516cf7a6ac306f29882e9c77558-Paper-Datasets_and_Benchmarks_Track.pdf.

Caijun Yan, Yang Zhou, Meixing Shi, Haoran Sun, Yichen Li, Yuxiang Cai, and Yankai Jiang. Reveal: A rubric-guided agent for explicit evidence sufficiency verificationin long-video question answering, 2026a. URL https://arxiv.org/abs/2608.08612.

Haiyang Yan, Hongyun Zhou, Peng Xu, Xiaoxue Feng, and Mengyi Liu. Symphony: A cognitively-inspired multi-agent system for long-video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24031–24041, June 2026b.

Rui Ye, Zhongwang Zhang, Kuan Li, Huifeng Yin, Zhengwei Tao, Yida Zhao, Liangcai Su, Liwen Zhang, Zile Qiao, Xinyu Wang, Pengjun Xie, Fei Huang, Jingren Zhou, Siheng Chen, and Yong Jiang. Agentfold: Long-horizon web agents with proactive context folding. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=IuZoTgsUws.

Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Yan Shu, Nicu Sebe, Ji-Rong Wen, and Zhicheng Dou. Videoexplorer: Advancing long-horizon video understanding via hierarchical orchestration. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 6350–6361, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3817705. URL https://doi.org/10.1145/3770855.3817705.

Lu Zhang, Tiancheng Zhao, Heting Ying, Yibo Ma, and Kyusong Lee. OmAgent: A multi-modal agent framework for complex video understanding with task divide-and-conquer. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10031–10045, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.559. URL https://aclanthology.org/2024.emnlp-main.559/.

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. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 89863–89895. Curran Associates, Inc., 2025. doi: 10.52202/085713-3005. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/8190b210e9808e54ee16263b673a847d-Paper-Conference.pdf.

Xintong Zhang, Xiaomeng Fan, Shilin Yan, Ekko He, Zicheng Liu, Zijian Zou, Guannan Zhang, Yuwei Wu, Zhi Gao, and Hongwei Xue. Adavdr: Adaptive tool use and reflection for video deep research, 2026. URL https://arxiv.org/abs/2608.25559.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701, June 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. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=To7Rs2wsTd.

A Implementation Details

A.1 Video Processing and Retrieval

We use Qwen3.8-27B for reasoning and visual observation. The initial question-level assessment receives 128 frames sampled uniformly over the video in chronological order, together with the available offline WhisperX transcript. Transcripts retain their original text, segment identifiers, and timestamps. Subsequent reasoning uses the accumulated observation text and source information; the initial frame set is not repeatedly appended to every agent call. Across the four datasets combined, our method uses an average of 204 video frames per question.

We construct a question-independent temporal index using OpenCLIP ViT-B/32. Index frames are sampled at 1 fps, independently of the 128-frame overview. Temporal boundaries are determined by changes in normalized OpenCLIP frame features and RGB color histograms, weighted by 0.75 and 0.25, respectively. Each change channel is normalized by its 95th percentile and clipped to $[0, 1]$. Boundary selection uses the larger of 0.35 and the 82nd percentile of the combined score, with a minimum spacing of 6 seconds; intervals longer than 45 seconds receive additional boundaries. Each interval is represented by the normalized mean of its frame embeddings. Importantly, the visual features and temporal segmentation are computed only once per video and cached for reuse across all questions and local search steps. At each subsequent planning step, only the current objective is encoded with OpenCLIP and matched against the cached interval embeddings by cosine similarity, with the top five intervals provided as navigation candidates. Consequently, repeated retrieval requires no additional video encoding or MLLM inference, introducing only the low overhead of text encoding and similarity ranking over cached features. The index remains fixed throughout inference, and the retrieved candidates serve only as navigation cues rather than restricting which intervals the agent may inspect.

The retrieval agent can inspect selected temporal regions or the full video at 0.1, 0.5, or 1 fps, access the transcript by time range or segment identifier, request a transcript-based search, or reuse existing observations. Visual requests are executed as bounded frame packets, while transcript reads return the original passages and source references. Transcript search uses model-based reference selection over the available text rather than a separate embedding score.

Each question has a shared ceiling of eight acquisition actions and seven local refinements. These limits are not reset for each new local objective. A local process returns on `deliver` or `limited`, or when its remaining execution allowance is exhausted. At the question-level action limit, the Controller selects a listed answer using the available observations. Decoding uses temperature 0 and a fixed random seed. Output limits are 512 tokens for objective generation, 4,096 for Local Audit and transcript search, and 8,192 for option assessment and transcript preprocessing. The remaining output limits and `top_p` use serving defaults. Thinking is enabled for final answer synthesis and disabled for the other stages.

A.2 Agent and Prompt Details

The Controller receives the original question, answer candidates, accumulated observations, and its task-level assessments, and decides whether to answer or issue a local retrieval objective. The Retrieval Agent receives that objective, relevant observations, navigation candidates, and local feedback, and chooses the next acquisition. Local Audit assesses whether the observations answer the objective (`deliver`), a concrete further observation is needed (`repair`), or the required capability is unavailable (`limited`). A `repair` decision continues locally, whereas handoff exposes observations and their sources without the intermediate retrieval trajectory.

The following abridged templates summarize each agent's inputs, outputs, and core decision rules. The Controller uses four separate calls: option assessment, sufficiency, objective generation, and final answer synthesis. Ordinary sufficiency requires `sufficient=true` and confidence at least 0.7.

A.2.1 CONTROLLER

Controller Four separate calls

C1 Option assessment

Input Question, answer options, current evidence states and source references.

Output One verdict and supporting references per option; an optional neutral observation need.

Assess each hypothesis formed by the question and an option, using only the current supplied observations. Options are hypotheses, not evidence. Return `SUPPORTED` when the observations collectively establish every material commitment; `CONTRADICTED` when they establish an incompatible fact; otherwise return `OPEN`. Missing support is not a contradiction. Match the question's scope, identity and temporal relations; for counts, establish distinct items without duplication.

Source alignment does not independently verify an observation or establish answer sufficiency. Preserve uncertainty and conflicts; do not extrapolate sampled facts into exhaustive counts or absence claims. Historical states cannot override current evidence. Cite the supporting claim and source references, and briefly state the covered commitment or remaining gap. Do not inspect the video or select a final answer in this call.

C2 Question-level sufficiency

Input Original question and options, video duration, cumulative observations and Controller context.

Output `sufficient`, `confidence`, `justification`, `reasoning`, and selected-option fields when sufficient.

Decide whether the accumulated observations are sufficient to answer the original question. If sufficient, identify the option and explain its support; otherwise, state the missing or uncertain information. Provide a confidence score in $[0, 1]$ reflecting the strength of that support. Judge from the supplied observations, without assuming access to raw video or guessing missing facts. The execution layer applies the sufficiency threshold described above.

C3 Local objective generation

Input Decision feedback, existing observations and unresolved gaps.

Output One `observation_need` containing `missing_proposition` and `affected_options`, or `null`.

Turn the feedback into one neutral, video-observable objective whose resolution could change an option assessment. The proposition must not contain option labels, answer judgments, or execution parameters such as timestamps, modality or frame rate. Retrieval planning owns those choices. Use existing facts and unresolved gaps; do not change option verdicts or invent observations. Return `null` if no useful observable objective remains.

C4 Final answer synthesis

Input Original question and options, video duration, accumulated observations and Controller context.

Output `selected_option`, `confidence`, `reasoning`, and `selected_option_text`.

Select a listed option and explain how the observations support it, citing relevant timestamps. At the question-level budget limit, select a listed option using the available observations and state any remaining uncertainty. Return the selected option, its text, confidence and concise reasoning as JSON.

A.2.2 RETRIEVAL AGENT

R1 Select the next acquisition

Input Active local objective, prior observations, navigation candidates, local attempts and Audit feedback.

Output One tool call with arguments, an explicit reuse action, or an unavailable action.

Choose one acquisition that addresses the current objective and remaining gap. Navigation candidates are hints, not observations. Keep the objective fixed while adapting what to inspect next. Do not answer the original question or invent evidence.

Available operations

- `observe_region`: inspect selected temporal intervals.
- `full_video_coverage`: inspect the full video when the objective requires that scope.
- `read_asr`: retrieve original transcript text by time, segment reference, or full-transcript scope.
- `search_asr`: obtain a transcript-based draft response with cited original passages and remaining uncertainty, for Local Audit to assess.
- `reuse`: cite existing observations for reassessment; `unavailable`: explain why an acquisition cannot currently help.

For visual calls, choose 0.1, 0.5 or 1 fps according to the needed temporal detail; the tool handles valid bounds, sampling and packetization. There is no mandatory sequence of frame rates or observation modalities. Return the selected tool and arguments as one JSON object.

Refinement and source use

For repeated or overlapping inspection, explain what the previous observations lack and what the new request can distinguish. Coverage alone does not establish that the objective was checked. Reuse supplied observations when adequate, and request additional context or a different scope when needed. An unavailable acquisition or unsuccessful search does not establish that the target is absent.

Choose visual or transcript access according to the gap; neither modality must always precede the other. Use returned transcript timestamps rather than inferring time from reference IDs. Utterance times need not be the times of the described events. When combining modalities, establish that they concern the same target and relation. Visual observation excludes raw ASR by default; already read passages may be supplied as optional guidance. A report echoing that guidance is not independent confirmation.

A.2.3 LOCAL AUDIT

A1 Resolve, refine, or return a limitation

Input Local objective, observation attempts, source materials, previous audits and available tools.

Output `action`, `answer`, supporting `references`, and an optional `gap`.

Assess the local objective using only the supplied observations. Check whether the observed identities, events, ordering, scope and relations support the response. State what is established and what remains unresolved. Preserve uncertainty and contradictions: source checks establish origin, not the correctness of a model description. Navigation ranking does not establish event order, and failure to locate a target does not prove absence.

A1 Resolve, refine, or return a limitation continued

Decision

- `deliver`: the observations answer the current objective. This does not establish that the original question or an answer option is proved.
- `repair`: a concrete further observation using the available tools could resolve the gap. Specify the missing fact; the Retrieval Agent chooses the acquisition parameters. An unlocated target or unexplored interval may still require retrieval.
- `limited`: the necessary capability, modality or essential task constraint is unavailable. Identify the limitation rather than proposing an unexecutable remedy.

Transcript and cross-modal checks

Audit transcript-derived draft answers against their cited original passages; the draft itself is not evidence. Request missing context or competing instances when needed, and cite original material when delivering. Transcript provenance does not guarantee recognition accuracy. Preserve unknown speaker identity, negations and later corrections. Shared time does not establish a shared event or speaker, and visual reports repeating supplied ASR do not provide independent confirmation. A single modality may suffice; do not require a second when the objective is already answered.

Return one JSON object with the decision, local response, supporting references and any remaining gap. Leave references empty when no supplied material supports the response. Do not select an answer option, plan execution parameters or claim an independent visual review.

B Experimental Details

B.1 Benchmark Protocols

Table 2 summarizes the multiple-choice evaluation settings. All four benchmarks use the same backbone and offline speech-processing pipeline.

Table 2: Benchmark evaluation protocols. Speech input is obtained through offline ASR.

Benchmark	Split / setting	Metric	Speech
LVBench	Standard	Accuracy	WhisperX ASR
LongVideoBench	Public validation	Accuracy	WhisperX ASR
Video-MME	Long split	Accuracy	WhisperX ASR
MLVU	Multiple-choice test	Accuracy	WhisperX ASR

Predictions are parsed into one of the question's supplied option labels and compared with the official answer by exact match; labels are unavailable to the inference agents and are loaded only after predictions are frozen. Responses are normalized to the original candidate list, allowing one text-only formatting correction without additional video acquisition or access to the correct answer.

B.2 Ablation Settings

Tables 3 and 4 specify the interventions. All variants within a comparison use the same backbone, initial observations, retrieval capabilities, per-question resource ceilings, and evaluation subset. Later observations may differ because the variants make different retrieval decisions.

Action-Level Return removes the local Audit/refinement loop and returns after one acquisition; the Controller remains free to launch another retrieval. Adding Local Status Feedback preserves the local assessment but returns control even when it recommends further refinement. Full Method instead uses this assessment to continue within the same objective before handoff.

Global Context $\rightarrow$ Local retains the same local objective and additionally exposes the question, options, and main-reasoning context to the local planner and Audit. Local Trace $\rightarrow$ Global retains the same local stopping rule but additionally returns the recorded planning, observation, and Audit

Table 3: Execution ablation. Local status denotes feedback from Local Audit.

Variant	Return after each action	Local status	Persistent local retrieval
Action-Level Return	✓	–	–
+ Local Status Feedback	✓	✓	–
Full Method	–	✓	✓

Table 4: Information-boundary ablation. All variants retain persistent local retrieval. Checkmarks indicate additional context exposure.

Variant	Full global context $\rightarrow$ local	Full local trace $\rightarrow$ global
Global Context $\rightarrow$ Local	✓	–
Local Trace $\rightarrow$ Global	–	✓
Full Method	–	–

trajectory at handoff. Full Method retains the restricted downward objective interface and the upward observation projection.

Retrieval action statistics. We additionally report the average number of acquisition actions per question in the execution-control ablation. Action-Level Return, Local Status Feedback, and Persistent Local Retrieval use 4.6, 4.3, and 3.7 acquisition actions on average, respectively. These results show that the performance gain of Persistent Local Retrieval is not explained by consuming more retrieval actions, but by organizing retrieval actions within a persistent local process before handoff.

Evaluation subsets. Each ablation study uses 597 randomly sampled questions: 180 from LVBench, 150 from LongVideoBench, 140 from Video-MME Long, and 127 from MLVU. All variants within each ablation are evaluated on the same question roster.

C Qualitative Case Studies

We examine two selected hard cases from runs using Qwen3.8-27B and 128 initial frames, consistent with the formal benchmark setting described in Appendix A.1. The cases illustrate how control and information boundaries shape local retrieval and subsequent global reasoning.

C.1 Establishing Event Order Before Answering

Figure 4 illustrates two dependent temporal relations: identifying the third performance in the video, then ordering the chair turns within it. An apparently relevant chair-turn observation cannot answer the question until the performance identity is established.

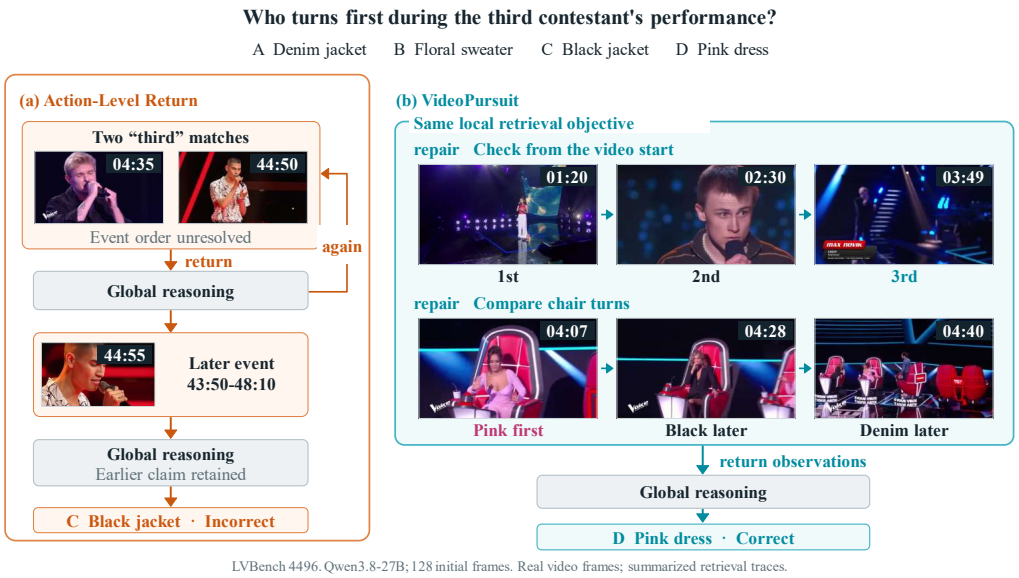

Figure 4: Event-order resolution on LVBench. Action-Level Return follows a later performance while retaining an earlier clothing interpretation, eventually selecting C. VideoPursuit keeps the objective local, establishes performance order, and then inspects the chair turns before returning observations, yielding the reference answer D. Real video frames accompany selected trace steps (Qwen3.8-27B, 128 initial frames); timestamps denote the displayed frames.

Local continuation. VideoPursuit retains one objective across four acquisitions and three `repair` decisions. After an uninformative candidate inspection, a coarse scan produces conflicting *third-contestant* reports. Inspecting 0–265 seconds establishes the target performance; inspecting 228–300 seconds then reports pink turning before black and denim. Audit returns `deliver`, and question-level reasoning supports D through the ordinary sufficiency exit.

Immediate-return contrast. The control instead inspects 2630–2890 seconds before resolving the global ordinal relation. Its option assessment retains a coarse *black T-shirt* to *black jacket* mapping despite conflicting later reports, ultimately answering C after six actions, including repeated observation and reuse. Continued searching therefore does not resolve the missing event identity in this trace. VideoPursuit's local process makes that gap determine the next acquisition: establish *which performance* before deciding *who turns first*.

C.2 Reconsidering a Local Interpretation at the Question Level

Case 2 MLVU: counting cookie-baking scenes

Question. In this video, how many instances are there of the 'baking cookies' action scene in total?

Options. A. 1; B. 6; C. 2; D. 0.

Reference answer: C. VideoPursuit: C (correct). Local Trace → Global: A (incorrect).

This example concerns the scope of an action category. Both runs obtain visual reports of an early ingredient-preparation segment and a later segment showing cookies being removed from an oven. Their local audits nevertheless narrow the count to active baking and exclude preparation. The information boundary determines whether this assembled local interpretation is part of the subsequent question-level input.

Shared local ambiguity. In VideoPursuit, the full-video observation describes *two distinct segments related to baking cookies*, with ingredient preparation around 46–56 seconds and an oven interaction around 250–260 seconds. A local refinement checks the oven segment and other candidate intervals. The terminal Audit concludes that there is *a total count of 1*, explaining that preparation *does not constitute actual baking*. The Local Trace → Global run similarly refines the oven segment and excludes the earlier mixing scene: *This is preparation, not the specific 'baking' actions requested.*

Different question-level inputs and decisions.

	VideoPursuit	**Local Trace → Global**
Returned material	Acquired observations and source information; the Audit's count of one is excluded.	Observations plus the recorded local trajectory, including the exclusion of preparation.
Question-level interpretation	Counts the two reported cookie-making segments: preparation and oven interaction.	Retains the narrower *active baking* definition and excludes the mixing segment.
Outcome	C: two scenes; matches the benchmark label.	A: one scene; differs from the benchmark label.

VideoPursuit's option assessment supports C using *two distinct segments . . . ingredient prep and actual baking*. The comparison contradicts C because there is *only one active baking scene; the other is ingredient mixing, not baking*. Both runs finish through ordinary sufficiency, with one local refinement each.

What the case illustrates. A useful set of observations can support a different interpretation when reconsidered in the context of the original question. By returning the acquired observations without carrying over the local Audit conclusion, VideoPursuit leaves question-level reasoning free to reassess the requested category and count. The key distinction is therefore not how much information is retained, but whether a local interpretation is propagated as part of the subsequent reasoning context.

The two runs differ in their sampled observations and intermediate trajectories. We use this case to illustrate how observations and local interpretations can play different roles after handoff to question-level reasoning.

Trace identifiers. Case 1 uses LVBench question `EwskdNETNx8_4496`; Case 2 uses MLVU question `Q595`. The displayed timings and descriptions are taken from recorded model observations, with segment-relative times converted to video time where necessary.