Agentic Active Omni-Modal Perception for Multi-Hop Audio-Visual Reasoning

Anonymous ACL submission

Abstract

Multi-hop audio-visual reasoning remains challenging for Omni-LLMs, as relevant evidence is often sparse, temporally dispersed, and distributed across both audio and visual streams. Existing benchmarks provide limited investigation of this setting, typically involving only a limited number of modalities, relevant temporal segments, or reasoning steps. In this work, we introduce MOV-Bench, a benchmark containing 519 carefully curated questions that require multi-hop reasoning over temporally dispersed audio-visual evidence. Evaluations on MOV-Bench reveal that current Omni-LLMs still struggle with multi-hop cross-modal reasoning. To address this challenge, we further propose AOP-Agent, an efficient agentic framework built on open-source Omni-LLMs for active omni-modal perception. By combining a hierarchical omni-modal memory with a collaborative observe-reflect-replan loop, AOP-Agent enables open-source Omni-LLMs to perform active perception without additional training or proprietary models. Experiments on MOV-Bench and OmniVideoBench demonstrate that AOP-Agent consistently improves reasoning performance, with particularly notable gains on long videos and reasoning-intensive questions.

1 Introduction

Recent end-to-end omni-modal large language models (Omni-LLMs) (Zhao et al., 2025; Li et al., 2025b; AI, 2025; Xu et al., 2025b; Team, 2026b; Yang et al., 2025; Wang et al., 2025a; Xu et al., 2026) have made substantial progress toward unified audio, visual, and textual understanding. Despite these advances, multi-hop audio-visual reasoning remains particularly challenging, since task-relevant evidence is often sparse, temporally dispersed, and distributed across both audio and visual streams. Answering such questions therefore requires cross-modal multi-hop reasoning over evidence appearing at different temporal locations.

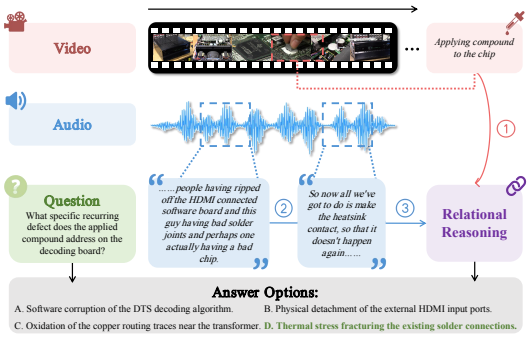


Figure 1: A representative example from MOV-Bench illustrating audio-visual multi-hop reasoning in the video. The question requires linking temporally separated visual evidence, which identifies the compound applied to the chip, with audio evidence describing the bad chip and the heatsink contact. The correct answer can only be obtained by jointly aligning these sparse audio-visual cues across time.

Existing video benchmarks have begun to include audio-visual understanding tasks (Li et al., 2025a; Hong et al., 2026; Wang et al., 2025b). However, their task composition is often broad and heterogeneous, making it difficult to isolate the specific capability of cross-modal multi-hop reasoning in videos. Even when existing questions involve multi-hop reasoning over videos, the number of reasoning steps or relevant temporal segments is often limited, and the decisive evidence may still be concentrated in a single modality (Chen et al., 2024a; Fu et al., 2024, 2026). As a result, existing evaluations provide limited insight into whether Omni-LLMs can jointly integrate audio and visual evidence distributed across multiple temporal segments to answer complex multi-hop questions.

To address this gap, we propose **MOV-Bench** (**M**ulti-hop **O**mni **V**ideo Benchmark), a benchmark for multi-hop audio-visual reasoning across multiple temporal segments(Figure 1). MOV-Bench contains 519 meticulously curated multiple-choice questions requiring models to integrate temporally

dispersed evidence across modalities and video segments. Based on MOV-Bench, we find that the key challenge for current Omni-LLMs lies not only in reasoning over audio-visual evidence, but also in effectively locating and acquiring the sparse, temporally dispersed evidence required for complex multi-hop reasoning.

This observation motivates a shift from passive video processing toward active and iterative perception. Unlike existing Omni-LLMs that passively encode entire videos, human viewers often selectively attend to relevant moments, revisit earlier evidence, and progressively refine their understanding. Such active observation is particularly important when decisive evidence is sparse or temporally dispersed (Wang et al., 2024a; Nguyen et al., 2024; Zhang et al., 2026; Qiu et al., 2026).

Recent agentic approaches have demonstrated the potential of active perception for audio-visual reasoning (Luo et al., 2025; Liu et al., 2025; Chen et al., 2026a). However, existing methods are often resource-intensive, either depending on costly proprietary models for multi-round decision making or requiring extensive training to equip open-source models with strong planning and reflection capabilities. This raises a key question: can open-source Omni-LLMs effectively achieve active perception without costly proprietary models or large-scale training?

To address this challenge, we propose **AOP-Agent** (**A**ctive **O**mni **P**erception Agent), an agentic framework for omni-modal reasoning under low-resource settings (Figure 2). AOP-Agent reduces the difficulty of active perception for open-source Omni-LLMs by organizing videos into a hierarchical omni-modal memory and decomposing complex audio-visual reasoning into collaborative observe-reflect-replan interactions. This design enables open-source Omni-LLMs to progressively localize and integrate sparse audio-visual evidence without additional training or proprietary models.

Experiments demonstrate that AOP-Agent consistently improves multi-hop omni-modal reasoning, especially on long videos and reasoning-intensive questions. AOP-Agent achieves substantial gains on both MOV-Bench and OmniVideoBench benchmark (Li et al., 2025a). Further comparisons with representative agentic video reasoning frameworks show that AOP-Agent enables more efficient active perception for open-source Omni-LLMs.

The main contributions of this work are summarized as follows:

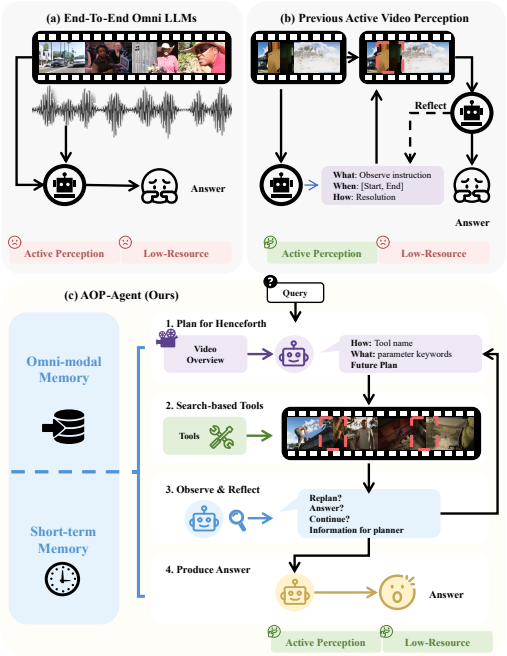


Figure 2: **Motivation of AOP-Agent.** Existing methods either passively process the full video or depend on high-resource settings such as extensive training or proprietary models. AOP-Agent enables Omni-LLMs to actively decide which video evidence to observe under low-resource settings. Short-term Memory denotes working memory and evidence memory in AOP-Agent.

- We introduce **MOV-Bench**, a benchmark for multi-hop audio-visual reasoning that requires integrating temporally dispersed evidence across multiple video segments and modalities.
- We propose **AOP-Agent**, a low-resource agentic framework that enables active perception for open-source Omni-LLMs through hierarchical omni-modal memory and collaborative observe-reflect-replan interactions, without additional training or proprietary models.
- Experiments on MOV-Bench and OmniVideoBench demonstrate that AOP-Agent consistently improves multi-hop omni-modal reasoning, with particularly notable gains on long videos and reasoning-intensive questions.

2 MOV-Bench

MOV-Bench is designed to evaluate whether Omni-LLMs can solve cross-modal multi-hop questions by locating and integrating temporally dispersed

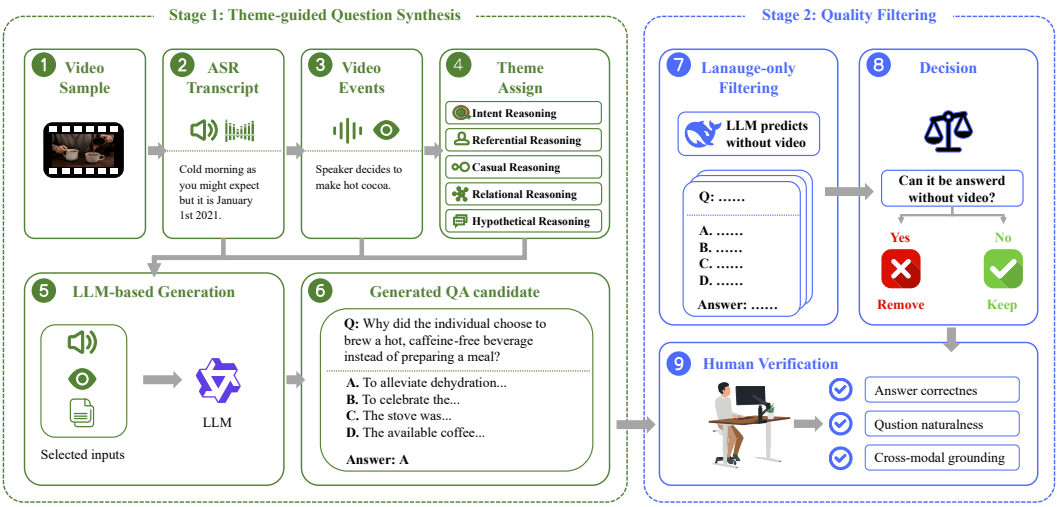


Figure 3: Overall pipeline for data collection, question synthesis, and quality filtering in MOV-Bench.

audio-visual evidence. Unlike conventional video QA tasks that can often be answered from a single clip or a single modality, MOV-Bench requires models to connect multiple pieces of evidence across different video segments and modalities. As shown in Figure 3, our construction pipeline consists of theme-guided question synthesis followed by automatic and human quality filtering.

2.1 Benchmark Design

MOV-Bench is built around three design principles.
Multi-hop temporal reasoning. Each question requires evidence from multiple video segments rather than a single moment. The reasoning chains involve 2, 3, or 4 hops, allowing MOV-Bench to evaluate models beyond short-range perception.
Cross-modal evidence integration. Questions are constructed such that neither the visual stream nor the audio stream alone is sufficient. Models must jointly interpret visual events, speech content, and audio cues to infer the answer.
Reasoning-oriented question themes. To cover diverse reasoning patterns, questions are organized around five themes: causal, referential, relational, hypothetical, and intent reasoning. These themes define distinct reasoning patterns and help ensure that MOV-Bench covers diverse forms of audio-visual inference.

2.2 Dataset Construction

We construct MOV-Bench from Fine-Video (Farré et al., 2024), which contains diverse videos with rich visual and auditory content. As shown in Figure 3, we adopt an LLM-based question generation pipeline. For each video, we first provide the LLM with ASR transcripts, video event descriptions, and a randomly selected reasoning theme. The LLM is then prompted to generate multiple-choice questions whose answers require evidence from both modalities and multiple temporal segments. To avoid unnatural or weakly grounded questions, the LLM is instructed to return empty content when the video does not contain sufficient evidence for a valid cross-modal multi-hop question.

2.2.1 Theme-guided Question Synthesis

We define five reasoning themes to guide question generation.
Causal Reasoning requires inferring the cause or consequence of an event by connecting audio and visual evidence.
Referential Reasoning requires resolving references in one modality using evidence from another, such as identifying a visual entity mentioned or implied in speech.
Relational Reasoning requires deriving social, spatial, or logical relationships between entities across modalities.
Hypothetical Reasoning requires inferring possible outcomes under altered conditions based on facts established in the video.
Intent Reasoning requires understanding a person's goal or motivation by jointly considering actions and spoken content.

2.2.2 Quality Filtering

We apply a two-stage filtering process to improve reliability. First, we perform a language-only filtering step: generated questions are given to LLMs

without video input, and questions that can be answered from linguistic bias or prior knowledge are removed. Second, human annotators verify the remaining samples for answer correctness, question naturalness, and consistency between the question, options, and supporting video evidence. This process ensures that MOV-Bench emphasizes genuine cross-modal multi-hop reasoning rather than shortcut-based answering. The statistics of MOV-Bench is in Appendix B.

3 AOP-Agent

MOV-Bench introduces a challenging setting for audio-visual reasoning, where answering questions requires integrating sparse audio-visual clues distributed across multiple temporal segments. A natural solution is active perception, where models iteratively identify and inspect relevant video segments during reasoning. However, existing approaches often rely on costly proprietary models or additional training to support such iterative observation.

Hence, we propose **AOP-Agent**, an efficient agentic framework for active omni-modal perception under low-resource settings. AOP-Agent combines hierarchical omni-modal memory construction with a multi-agent observe-reflect-replan loop, enabling open-source Omni-LLMs to progressively refine their understanding through iterative audio-visual observation without additional training or proprietary components. Detailed cases are provided in Appendix E.

3.1 Overview of AOP-Agent

As illustrated in Figure 4, AOP-Agent performs audio-visual reasoning through a multi-agent observe-reflect-replan process over hierarchical omni-modal memory. Given a long video V and a natural language question Q, AOP-Agent first converts the raw video stream into a structured omni-modal memory $\mathcal{M}_V$. The agents then progressively observe, reflect, and replan over the memory to refine their understanding of the video before generating the final answer A.

Hierarchical Omni-modal Memory. AOP-Agent organizes the video into a hierarchical memory with multiple temporal and semantic granularities, including global video summaries, segment-level descriptions, audio-visual keypoints, and retrieval-oriented keywords. This hierarchical structure supports efficient localization and coarse-to-fine observation.

Active Omni-modal Perception. Based on the hierarchical memory $\mathcal{M}_V$, AOP-Agent performs iterative audio-visual perception through a multi-agent observe-reflect-replan loop. Specifically, the planner agent A_P determines the next observation target according to the current reasoning state, specifying the video regions that should be further inspected. The observation tools W_T then follow the planner's instruction to retrieve the corresponding mid-level segments. Based on these returned segments, the reflector agent A_R evaluates whether the current information is sufficient or further observation is needed. Once sufficient information has been accumulated across iterations, the reasoner agent A_A generates the final answer. More details are provided in Appendix D.3.

3.2 Hierarchical Omni-modal Memory Construction

Audio-visual reasoning in long videos is limited by context capacity and attention dilution. Instead of directly processing the entire video stream, AOP-Agent organizes the video into a hierarchical omni-modal memory that supports coarse-to-fine evidence localization and observation.

Specifically, we apply a video segmentation strategy based on the timestamps of ASR results. The original video is split into N_{fine} fine-grained clips to preserve local audio-visual details as: $\{v_i^{\text{fine}} = (b_i^{\text{fine}}, e_i^{\text{fine}})\}_{i=1}^{N_{\text{fine}}}$, where b_i^{fine} and e_i^{fine} denote the start and end timestamps of the i-th fine-grained clip v_i. Adjacent clips are then merged into N_{mid} mid-level semantic segments, each no longer than 30 seconds, and compressed at a lower resolution as: $\{v_i^{\text{mid}} = (b_i^{\text{mid}}, e_i^{\text{mid}})\}_{i=1}^{N_{\text{mid}}}$. For each mid-level segment v_i^{mid}, we use an Omni-LLM to generate structured semantic information, including visual keypoints K_i^v that characterize events depicted in the visual content of the segment, audio keypoints K_i^a that capture events present in the audio stream of the segment, retrieval keywords K_i^w that concisely summarize the segment, and a segment-level description $D_{V,i}$ that provides a sentence-level summary of the segment. The global video description D_V is then synthesized from all segment-level descriptions: $D_V = \text{LLM}\left(\{D_{V,i}\}_{i=1}^{N_{\text{mid}}}\right)$. Examples of the structured semantic information are provided in Appendix D.1.

The hierarchical memory provides both global video context and fine-grained audio-visual observations, enabling AOP-Agent to adaptively se-

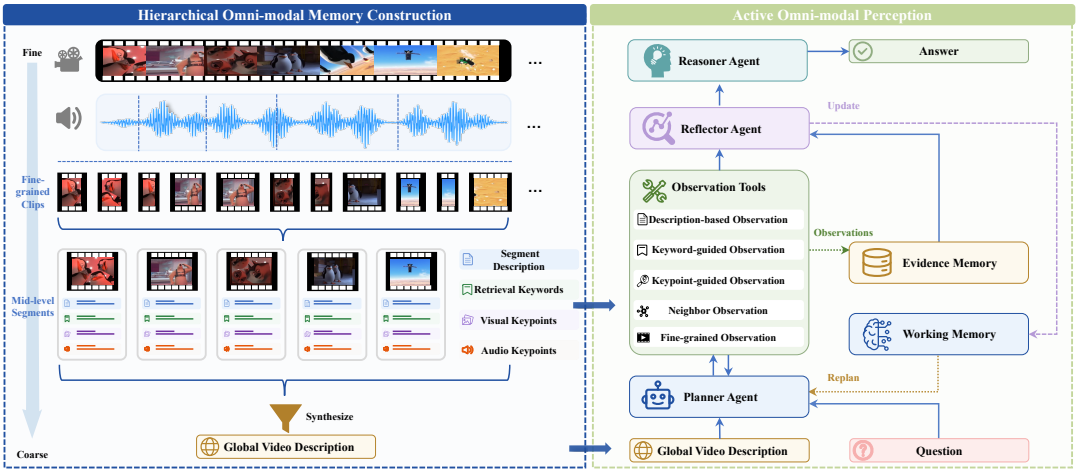


Figure 4: **Framework of AOP-Agent.** AOP-Agent reformulates omni-modal reasoning as active perception through hierarchical omni-modal memory and an observe-reflect-replan loop.

lect observation granularity, and efficiently inspect relevant video content during iterative perception. More implementation details of the video segmentation strategy are provided in Appendix D.4.

3.3 Active Omni-modal Perception

After constructing the hierarchical omni-modal memory, AOP-Agent performs active perception through a multi-agent observe-reflect-replan loop. Instead of uniformly processing the entire video at fine granularity, AOP-Agent dynamically determines where to observe and how finely to inspect according to the current reasoning state. To support flexible perception over long videos, AOP-Agent employs a set of complementary observation tools.

Description-based Observation Tool. This tool locates semantically relevant mid-level segments by matching planner queries against segment-level descriptions $\{D_{V,i}\}_{i=1}^{N_{\text{mid}}}$. It is particularly effective for observing high-level events, intentions, and relational information.

Keyword-guided Observation Tool. This tool identifies segments containing explicit entities, actions, or spoken content through keyword-based matching over semantic keywords $\{K_i^w\}_{i=1}^{N_{\text{mid}}}$.

Keypoint-guided Observation Tool. This tool matches visual keypoints $\{K_i^v\}_{i=1}^{N_{\text{mid}}}$ and audio keypoints $\{K_i^a\}_{i=1}^{N_{\text{mid}}}$ to identify cross-modal clues related to the current reasoning state.

Neighbor Observation Tool. Given an observed segment, this tool expands perception to temporally adjacent segments, allowing AOP-Agent to recover related events that occur before or after the current observation target.

Fine-grained Observation Tool. Given a selected mid-level segment, this tool further inspects its corresponding fine-grained clips for detailed local verification when coarse-grained observations are insufficient.

Together, these observation tools provide complementary ways to navigate the hierarchical memory, enabling AOP-Agent to progressively localize relevant video regions and adaptively switch between coarse-grained and fine-grained perception during iterative reasoning.

During inference, the Planner Agent A_P determines the next observation target according to the question Q, the global video description D_V, and the current Working Memory M_W. The selected observation tool then retrieves the corresponding mid-level segments for further inspection and stores them in the Evidence Memory M_E. The Reflector Agent A_R evaluates whether the current observations are sufficient for reasoning and provides feedback for subsequent replanning if necessary. Once sufficient information has been accumulated across iterations, the Reasoner Agent A_A generates the final answer based on the collected observations.

Through the observe-reflect-replan loop, AOP-Agent progressively narrows its observation scope and performs adaptive multi-grained perception for reliable multi-hop cross-modal reasoning over long videos. By combining hierarchical omni-modal memory with collaborative multi-agent decomposition, AOP-Agent enabling open-source Omni-LLMs to perform effective iterative video reasoning without costly proprietary models or additional training. Details of the observation tools and the complete agentic framework are provided in Ap-

pendix D.2 and Appendix D.3, respectively.

4 Experiment

In this section, we present experimental results on MOV-Bench and OmniVideoBench to evaluate the capabilities of current Omni-LLMs and the effectiveness of AOP-Agent for cross-modal multi-hop video reasoning. Specifically, we study the following research questions:

RQ1: What limitations of current Omni-LLMs are exposed by MOV-Bench in cross-modal multi-hop video reasoning?

RQ2: Can AOP-Agent improve cross-modal multi-hop reasoning across different Omni-LLMs and benchmarks?

RQ3: How does AOP-Agent differ from existing agentic video reasoning frameworks under open-source Omni-LLM settings?

RQ4: How do planning, reflection, model assignment, and observation rounds contribute to the performance of AOP-Agent?

4.1 Experimental Setup

4.1.1 Benchmarks

We conduct experiments on both MOV-Bench and OmniVideoBench. MOV-Bench evaluates cross-modal multi-hop reasoning over temporally dispersed audio-visual evidence, while OmniVideoBench serves as a more general audio-visual understanding benchmark for evaluating the generalization ability of Omni-LLMs and AOP-Agent. Additional benchmark details are provided in Appendix C.3.

4.1.2 Evaluated Models and Baselines

We evaluate representative Omni-LLMs including Qwen3-Omni-Instruct, Qwen3-Omni-Thinking, Qwen3-Omni-Captioner (Xu et al., 2025b), Qwen2.5-Omni-7B, Qwen2.5-Omni-3B (Xu et al., 2025a), Ming-Lite-Omni-1.5 (AI, 2025), Baichuan-Omni-1.5 (Li et al., 2025b), and MiniCPM-o-4.5 (Cui et al., 2026).

For agentic baselines, we evaluate OmniAgent (Tao et al., 2026) and ActiveVideoPerception (Wang et al., 2025e), which represent recent frameworks for active perception and agentic audio-visual reasoning. To ensure fair comparison under open-source settings, both frameworks are re-implemented using Qwen3-Omni-Instruct as the backbone model. Additional implementation details are provided in Appendix C.1.

4.1.3 Implementation of AOP-Agent

To evaluate AOP-Agent across different Omni-LLMs, we implement AOP-Agent using Qwen3-Omni-Thinking, Qwen3-Omni-Instruct, and Qwen2.5-Omni-7B as backbone models, where all agents share the same backbone unless otherwise specified.

For hierarchical memory construction, we use Qwen3-Omni-Instruct to generate segment-level audio-visual descriptions, keypoints, and global video summaries. We adopt FunASR (Gao et al., 2023) for ASR-based temporal segmentation.

4.2 Results on MOV-Bench (RQ1)

Table 1 presents the main results on MOV-Bench. Overall, current open-source Omni-LLMs achieve relatively low performance, suggesting that cross-modal multi-hop reasoning remains highly challenging. For example, Qwen3-Omni-Instruct achieves only 52.79% overall accuracy under direct inference, highlighting the difficulty of cross-modal multi-hop reasoning over long videos.

Long videos substantially increase reasoning difficulty. Across different Omni-LLMs, performance consistently decreases as video length increases. For example, Qwen3-Omni-Instruct drops from 62.43% on short videos to 45.50% on long videos subset. Similar trends are observed across other models, suggesting that current Omni-LLMs struggle to preserve and retrieve sparse evidence over long temporal spans.

Higher-hop questions remain challenging for current Omni-LLMs. Most Omni-LLMs achieve lower performance on 4-hop questions compared with their overall accuracy. For example, Qwen3-Omni-Thinking drops from 55.49% overall accuracy to 44.19% on 4-hop questions, while Ming-Lite-Omni-1.5 decreases from 49.52% to 41.86%. These results suggest that current Omni-LLMs still struggle to connect sparse audio-visual clues distributed across multiple temporal segments.

MOV-Bench reveals a key bottleneck in current Omni-LLMs. Taken together, these results suggest that a significant limitation of current Omni-LLMs lies not only in reasoning itself, but also in effectively locating, preserving, and integrating sparse cross-modal evidence distributed across long videos.

4.3 Results of AOP-Agent (RQ2)

Table 1 shows that AOP-Agent consistently improves direct-inference baselines across dif-

Model	Causal	Ref	Hypo	Rel	Intent	Long	Medium	Short	2hop	3hop	4hop	Overall
Baseline (Direct Inference)												
Qwen3-Omni-Instruct	48.35	**59.63**	35.00	54.33	55.00	45.50	49.65	62.43	54.92	49.28	51.16	52.79
Qwen3-Omni-Thinking	56.04	57.14	35.00	60.63	58.75	48.15	51.77	65.61	57.29	58.70	44.19	55.49
Qwen3-Omni-Captioner	59.34	57.76	30.00	51.97	57.50	46.56	53.19	60.32	52.88	54.35	53.49	53.37
Qwen2.5-Omni-7B	41.76	52.80	25.00	55.12	42.50	37.57	45.39	56.61	49.15	45.65	39.53	46.63
Qwen2.5-Omni-3B	42.86	45.96	38.33	44.09	43.75	35.98	43.97	51.32	44.41	46.38	37.21	43.74
Ming-Lite-Omni-1.5	46.15	55.90	35.00	50.39	50.00	44.44	47.52	56.08	49.83	53.62	41.86	49.52
Baichuan-Omni-1.5	41.76	47.83	30.00	44.09	47.50	37.57	39.72	52.91	45.76	45.65	33.72	43.74
MiniCPM-o 4.5	54.95	58.39	40.00	54.33	48.75	43.39	53.90	62.43	56.27	50.00	47.67	53.18
Other Agentic Framework												
OmniAgent	45.05	38.51	28.33	35.43	42.50	33.86	39.72	41.80	36.61	45.65	32.56	38.34
Active Video Perception	38.46	32.30	34.48	30.16	46.25	32.80	38.85	35.11	33.79	39.42	33.72	35.27
AOP-Agent (Ours)												
Qwen3-Omni-Instruct	**71.43**	58.39	**50.00**	62.99	**70.00**	**60.85**	62.41	64.55	**64.75**	61.59	56.98	**62.62**
Qwen3-Omni-Thinking	70.33	55.28	48.33	**68.50**	67.50	55.03	**63.12**	**68.78**	60.68	**67.39**	**59.30**	62.24
Qwen2.5-Omni-7B	50.55	47.83	41.67	51.18	38.75	41.27	46.81	52.91	48.14	47.10	43.02	47.01

Table 1: Results of different models and frameworks. The table reports accuracy on videos across five reasoning types, three duration ranges and three hop types. Boldface highlights the best performace within each column.

Model	Long	Medium	Short	Reason	Overall
Baseline (Direct Inference)					
Qwen3-Omni-Instruct	28.45	26.27	38.21	31.09	29.30
Qwen3-Omni-Thinking	31.90	**37.29**	37.40	35.57	31.90
Qwen2.5-Omni-7B	30.17	27.12	34.15	30.53	29.00
Other Video Agent Frameworks					
OmniAgent	28.45	34.75	31.71	31.65	27.30
Active Video Perception	30.17	22.88	36.89	30.06	26.90
AOP-Agent (Ours)					
Qwen3-Omni-Instruct	**40.52**	33.05	**39.84**	37.82	**33.20**
Qwen3-Omni-Thinking	**40.52**	35.59	**39.84**	**38.66**	32.90
Qwen2.5-Omni-7B	31.90	33.05	34.15	33.05	31.10

Table 2: Comparison of the baselines and AOP-Agent on OmniVideoBench. Boldface highlights the best performace within each column.

ferent Omni-LLM backbones on MOV-Bench, including Qwen3-Omni-Instruct, Qwen3-Omni-Thinking, and Qwen2.5-Omni-7B. The gains are especially pronounced on long videos and reasoning-intensive questions. For example, with Qwen3-Omni-Instruct, AOP-Agent improves long-video accuracy from 45.50% to 60.85%. Table 2 further shows that AOP-Agent generalizes to OmniVideoBench, where it also improves performance on long-video and reasoning-related subsets.

These results suggest that active perception is particularly beneficial for long-video and reasoning-intensive scenarios, where decisive evidence is often sparse and temporally dispersed. Under direct full-video inference, Omni-LLMs must preserve and retrieve sparse question-relevant evidence from long videos filled with irrelevant context, making multi-hop reasoning highly challenging. In contrast, AOP-Agent progressively narrows the observation space through iterative evidence acquisition and reflection, allowing the model to focus on question-relevant segments and perform more reliable cross-modal evidence integration for multi-hop reasoning.

4.4 Comparison with Other Agentic Frameworks (RQ3)

We further compare AOP-Agent with OmniAgent and ActiveVideoPerception using Qwen3-Omni-Instruct as the backbone model.

As shown in Table 1, existing agentic frameworks perform substantially worse than direct inference on MOV-Bench. For example, OmniAgent and ActiveVideoPerception achieve only 38.34% and 35.27% overall accuracy, both below direct Qwen3-Omni-Instruct inference. Similar trends are also observed on OmniVideoBench (Table 2).

These results suggest that existing agentic video reasoning frameworks do not naturally transfer to open-source Omni-LLMs. Prior methods are typically designed around powerful proprietary models with strong planning and long-horizon reasoning capabilities. As a result, directly introducing complex iterative observation may instead amplify error accumulation for weaker open-source models under sparse and temporally dispersed evidence.

In contrast, AOP-Agent consistently outperforms both direct inference and existing agentic frameworks across benchmarks. The key difference is that AOP-Agent explicitly reduces the difficulty of active perception for open-source Omni-LLMs. Its hierarchical omni-modal memory makes sparse evidence easier to localize and retrieve, while the multi-agent collaboration framework decomposes

Component	MOV Long	OVB Long
Reasoner (Baseline)	45.50	28.45
+ Planner	50.26	34.48
+ Planner + Reflector (Full)	**60.85**	**40.52**

Table 3: Ablation Study on long-video subset of OmniVideoBench (OVB) and MOV-Bench (MOV).

Planner & Reflector	Reasoner	MOV Long	OVB Long
Qwen2.5-Omni-7B	Qwen2.5-Omni-7B	41.27	31.90
Qwen3-Omni-30B	Qwen2.5-Omni-7B	48.68	33.62
Qwen2.5-Omni-7B	Qwen3-Omni-30B	58.20	35.34
Qwen3-Omni-30B	Qwen3-Omni-30B	**60.85**	**40.52**

Table 4: Performance comparison of different model assignments on MOV-Bench (MOV) and OmniVideoBench (OVB).

active perception into simpler subtasks. Together, these designs make active perception substantially more tractable for open-source Omni-LLMs.

4.5 Ablation Study (RQ4)

We conduct ablation experiments on the long-video subsets of MOV-Bench and OmniVideoBench, where sparse and temporally dispersed evidence makes active perception particularly important, to analyze the role of planning, reflection, model assignment, and observation rounds in AOP-Agent.

4.5.1 Effect of AOP-Agent Components

Table 3 presents the contribution of different components in AOP-Agent on long-video subset from MOV-Bench and OmniVideoBench.

Adding the Planner consistently improves performance over direct reasoning, suggesting that actively selecting observation targets is beneficial for long-video omni-modal reasoning. Further introducing the Reflector leads to the strongest performance, indicating that iterative reflection helps stabilize active perception by verifying current observations and guiding subsequent replanning when necessary.

4.5.2 Effect of Model Assignment

Table 4 analyzes different model assignments within AOP-Agent. Stronger planners and reflectors consistently improve performance even when the final reasoner remains unchanged, while weaker planners and reflectors noticeably limit overall performance despite using a stronger reasoner.

These results suggest that the effectiveness of active perception is largely determined before final reasoning begins. High-quality planning and reflection lead to more reliable observation trajectories and evidence selection, ultimately determining whether the reasoner receives useful evidence for complex multi-hop reasoning.

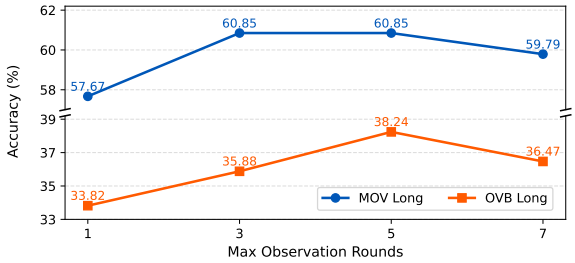


Figure 5: Performance of AOP-Agent based on Qwen3-Omni-Instruct on long-video subsets under different maximum observation round limits.

4.5.3 Effect of Observation Rounds

Figure 5 shows that increasing the maximum observation rounds from 1 to 3 improves performance, indicating that iterative observation helps AOP-Agent recover missed evidence and refine its evidence memory. However, further increasing the round limit leads to saturation or slight degradation, suggesting that excessive observation can introduce noisy or irrelevant segments. Overall, the results suggest that active observation should remain bounded: AOP-Agent performs best when it acquires sufficient evidence without accumulating excessive irrelevant context.

5 Conclusion

We introduce MOV-Bench, a benchmark for evaluating multi-hop audio-visual reasoning in videos, and AOP-Agent, an agentic framework for active omni-modal perception under low-resource settings. Results on MOV-Bench show that current Omni-LLMs are limited not only by reasoning over audio-visual evidence, but also by their ability to locate and acquire sparse evidence distributed across audio and vision streams. By combining hierarchical omni-modal memory with an observe-reflect-replan loop, AOP-Agent enables models to selectively inspect question-relevant evidence without additional training or proprietary models. Extensive experiments on MOV-Bench and OmniVideoBench demonstrate consistent improvements across multiple Omni-LLMs, highlighting active perception as a promising paradigm for scalable and effective long-video omni-modal reasoning.

Limitations

Although MOV-Bench and AOP-Agent advance omni-modal video understanding and reasoning, several limitations remain. First, due to the cost of annotation and verification, the scale of MOV-Bench remains limited. Although human verification is included in the construction pipeline, synthesized questions may still be less diverse and natural than fully human-labeled data. Second, although AOP-Agent improves performance on both MOV-Bench and OmniVideoBench, its multi-round observe-reflect-replan loop introduces additional inference overhead. Third, as illustrated by the failure case in Figure 8, hallucination and error propagation remain critical challenges in the omni-modal reasonings. Although the reflector can partially mitigate error propagation, the overall accuracy is still constrained by the quality of the hierarchical omni-modal memory base and the reliability of the underlying OmniLLM. Finally, the proposed omni-modal memory base is not streaming-friendly, as it relies on a multi-level offline construction process. This may limit its applicability to real-time scenarios such as smart glasses and live streaming.

Ethical Considerations

MOV-Bench is constructed from Fine-Video (Farré et al., 2024), a publicly available dataset released for academic use. We comply with the license terms of Fine-Video, use the videos solely for research purposes, and do not redistribute any raw video content. Our benchmark construction process derives information from the original videos and does not intentionally introduce any new personally identifiable information. Due to the scale of MOV-Bench, the authors served as human annotators. Human annotators were involved in verifying the quality, correctness, and evidence alignment of the generated samples. We used large language models as writing assistants to edit and polish the paper. The authors reviewed and revised the generated text and remain fully responsible for the content, claims, and integrity of the work. Overall, we believe that this work introduces minimal additional ethical or privacy risks beyond those already associated with the underlying Fine-Video dataset.

References

Inclusion AI. 2025. Ming-omni: A unified multimodal model for perception and generation. *Preprint*, arXiv:2506.09344.

Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. 2025. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. *Preprint*, arXiv:2503.10200.

Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. 2024a. Cg-bench: Clue-grounded question answering benchmark for long video understanding. *Preprint*, arXiv:2412.12075.

Houlun Chen, Xin Wang, Guangyao Li, Yuwei Zhou, Yihan Chen, Jia Jia, and Wenwu Zhu. 2026a. Think with grounding: Curriculum reinforced reasoning with video grounding for long video understanding. *Preprint*, arXiv:2602.18702.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024b. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Qian Chen, Jinlan Fu, Changsong Li, See-Kiong Ng, and Xipeng Qiu. 2026b. Futureomni: Evaluating future forecasting from omni-modal context for multimodal llms. *Preprint*, arXiv:2601.13836.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, and 17 others. 2026. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

Sunqi Fan, Meng-Hao Guo, and Shuojin Yang. 2025. Agentic keyframe search for video question answering. *Preprint*, arXiv:2503.16032.

Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. 2024. Finevideo. `https://huggingface.co/datasets/HuggingFaceFV/finevideo`.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, and 1 others. 2026. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, and Shiliang Zhang. 2023. Funasr: A fundamental end-to-end speech recognition toolkit. In *INTERSPEECH*.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. 2026. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *Preprint*, arXiv:2502.04326.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Jiafu Tang, Zhenghao Song, Dingling Zhang, Ying He, Haoxiang Liu, Yuxuan Wang, Qiufeng Wang, Zhenhe Wu, Jiehui Luo, Zhiyu Pan, Weihao Xie, Chenchen Zhang, and 23 others. 2025a. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. *Preprint*, arXiv:2510.10689.

Xiaoxi Li, Wenxiang Jiao, Jiarui Jin, Shijian Wang, Guanting Dong, Jiajie Jin, Hao Wang, Yinuo Wang, Ji-Rong Wen, Yuan Lu, and Zhicheng Dou. 2026. Omnigaia: Towards native omni-modal ai agents. *Preprint*, arXiv:2602.22897.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, and 1 others. 2025b. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*.

Chengwen Liu, Xiaomin Yu, Zhuoyue Chang, Zhe Huang, Shuo Zhang, Heng Lian, Kunyi Wang, Rui Xu, Sen Hu, Jianheng Hou, and 1 others. 2026. Watching, reasoning, and searching: A video deep research benchmark on open web for agentic video reasoning. *arXiv preprint arXiv:2601.06943*.

Runtao Liu, Ziyi Liu, Jiaqi Tang, Yue Ma, Renjie Pi, Jipeng Zhang, and Qifeng Chen. 2025. Longvideoagent: Multi-agent reasoning with long videos. *Preprint*, arXiv:2512.20618.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. 2025. Videorag: Visually-aligned retrieval-augmented long video comprehension. *Preprint*, arXiv:2411.13093.

Thong Nguyen, Yi Bin, Junbin Xiao, Leigang Qu, Yicong Li, Jay Zhangjie Wu, Cong-Duy Nguyen, See-Kiong Ng, and Anh Tuan Luu. 2024. Video-language understanding: A survey from model architecture, model training, and data perspectives. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3636–3657, Bangkok, Thailand. Association for Computational Linguistics.

Jongwoo Park, Kanchana Ranasinghe, Kumara Kahatapitiya, Wonjeong Ryu, Donghyun Kim, and Michael S. Ryoo. 2026. Too many frames, not all useful: Efficient strategies for long-form video qa. *Preprint*, arXiv:2406.09396.

Jihao Qiu, Lingxi Xie, Xinyue Huo, Qi Tian, and Qixiang Ye. 2026. Longvideo-r1: Smart navigation for low-cost long video understanding. *Preprint*, arXiv:2602.20913.

Keda Tao, Wenjie Du, Bohan Yu, Weiqiang Wang, Jian Liu, and Huan Wang. 2026. Active perception agent for omnimodal audio-video understanding. *Preprint*, arXiv:2512.23646.

Qwen Team. 2026a. Qwen3.5: Accelerating productivity with native multimodal agents.

Qwen Team. 2026b. Qwen3.5-omni technical report. *Preprint*, arXiv:2604.15804.

Siyin Wang, Jinlan Fu, Feihong Liu, Xinzhe He, Huangxuan Wu, Junhao Shi, Kexin Huang, Zhaoye Fei, Jingjing Gong, Zuxuan Wu, Yugang Jiang, See-Kiong Ng, Tat-Seng Chua, and Xipeng Qiu. 2025a. Roboomni: Proactive robot manipulation in omni-modal context. *arXiv preprint arXiv:2510.23763*.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. 2024a. Lvbench: An extreme long video understanding benchmark. *Preprint*, arXiv:2406.08035.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024b. Videoagent: Long-form video understanding with large language model as agent. *Preprint*, arXiv:2403.10517.

Yueqian Wang, Xiaojun Meng, Yifan Wang, Huishuai Zhang, and Dongyan Zhao. 2025b. Proactivevideoqa: A comprehensive benchmark evaluating proactive interactions in video large language models. *Preprint*, arXiv:2507.09313.

Yuxuan Wang, Yueqian Wang, Bo Chen, Tong Wu, Dongyan Zhao, and Zilong Zheng. 2025c. Omnimmi: A comprehensive multi-modal interaction benchmark in streaming video contexts. *Preprint*, arXiv:2503.22952.

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

Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S. Ryoo, and Juan Carlos Niebles. 2025e. Active video perception: Iterative evidence seeking for agentic long video understanding. *Preprint*, arXiv:2512.05774.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025a. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Ke Xu, Yuhao Wang, and Yu Wang. 2026. From reactive to proactive: Assessing the proactivity of voice agents via provoice-bench. *Preprint*, arXiv:2604.15037.

Bufang Yang, Lilin Xu, Liekang Zeng, Yunqi Guo, Siyang Jiang, Wenrui Lu, Kaiwei Liu, Hancheng Xiang, Xiaofan Jiang, Guoliang Xing, and Zhenyu Yan. 2025. Proagent: Harnessing on-demand sensory contexts for proactive llm agent systems. *Preprint*, arXiv:2512.06721.

Ce Zhang, Yan-Bo Lin, Ziyang Wang, Mohit Bansal, and Gedas Bertasius. 2026. Silvr: A simple language-based video reasoning framework. *Preprint*, arXiv:2505.24869.

Jiaxing Zhao, Qize Yang, Yixing Peng, Detao Bai, Shimin Yao, Boyuan Sun, Xiang Chen, Shenghao Fu, Xihan Wei, Liefeng Bo, and 1 others. 2025. Humanomni: A large vision-speech language model for human-centric video understanding. *arXiv preprint arXiv:2501.15111*.

Yiyang Zhou, Yangfan He, Yaofeng Su, Siwei Han, Joel Jang, Gedas Bertasius, Mohit Bansal, and Huaxiu Yao. 2025a. Reagent-v: A reward-driven multi-agent framework for video understanding. *Preprint*, arXiv:2506.01300.

Ziwei Zhou, Rui Wang, and Zuxuan Wu. 2025b. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *Preprint*, arXiv:2505.17862.

Yifan Zhu, Xinyu Mu, Tao Feng, Zhonghong Ou, Yuning Gong, and Haoran Luo. 2026. Omnirag-agent: Agentic omnimodal reasoning for low-resource long audio-video question answering. *Preprint*, arXiv:2602.03707.

A Related Work

Multimodal Benchmarks. Benchmarks for multimodal large language models have evolved from basic video perception (Wang et al., 2024b) to more comprehensive audio-visual understanding and agentic reasoning (Wang et al., 2025e; Tao et al., 2026; Zhu et al., 2026). Recent omni-modal benchmarks, including OmniVideoBench (Li et al., 2025a), Daily-Omni (Zhou et al., 2025b), and WorldSense (Hong et al., 2026), evaluate video understanding with both visual and auditory inputs. Other benchmarks, such as Video-MME (Fu et al., 2024), FutureOmni (Chen et al., 2026b), and Video-DR (Liu et al., 2026), further extend evaluation to comprehensive video understanding, future forecasting, and open-world video reasoning. OmniMMI (Wang et al., 2025c) studies multimodal interaction, while OmniGAIA (Li et al., 2026) evaluates tool-integrated omni-modal agents. Different from these benchmarks, MOV-Bench focuses on multi-hop audio-visual reasoning, where models must integrate temporally dispersed evidence from both audio and visual streams.

Agentic Frameworks for Video Understanding. To mitigate long-context limitations, prior work has explored caption-based and retrieval-based video representations (Fan et al., 2025; Wang et al., 2025d; Park et al., 2026). These methods reduce input length by converting video segments into textual descriptions, but they usually rely on static representations and lack adaptive evidence acquisition. Recent agentic systems introduce planning, grounding, and reflection into video reasoning (Chen et al., 2026a), including LVAgent (Chen et al., 2025), ReAgent-V (Zhou et al., 2025a), Active Video Perception (Wang et al., 2025e), and OmniAgent (Tao et al., 2026). However, many of these methods depend on costly proprietary models or extensive training, making them resource-intensive. AOP-Agent instead adopts an agentic active perception framework, enabling models to iteratively decide what evidence to observe through planning and reflection with low resource requirements.

B Statistics of MOV-Bench

With videos from Fine-Video, their ASR-transcription and visual events within video events is provided in Fine-Video and audio events within video events is constructed with Qwen3-Omni-Instruct. We initially generate 2,240 candidate samples using Qwen3.5-Plus (Team, 2026a). After preliminary multiple-choice validation with DeepSeek-V4-Flash (DeepSeek-AI, 2026), 802 samples are retained. Following final human expert verification, MOV-Bench contains 519 high-quality multiple-choice questions.

As shown in Figure 6, MOV-Bench covers 15 real-world video categories. The average video duration is 240.26 seconds, and long videos exceeding 5 minutes account for 36.42% of the benchmark. The average question length is 17.02 words, while the average answer length is 7.87 words. To ensure multi-hop reasoning difficulty, MOV-Bench includes 295 two-hop questions, 138 three-hop questions, and 86 four-hop questions, with an average of 2.60 reasoning hops per sample.

C Experiment Details

C.1 Baselines

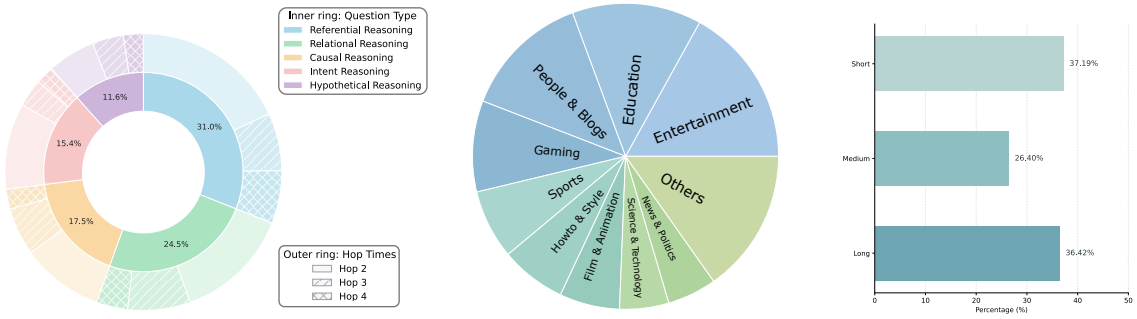
AOP-Agent follows an agentic active perception paradigm. To compare AOP-Agent with representative agentic video understanding frameworks, we select OmniAgent and ActiveVideoPerception as baselines. For ActiveVideoPerception, we replace the original Gemini backbone with Qwen3-Omni-Instruct. For OmniAgent, whose audio and video tools may use different Omni-LLMs as backbone models, we replace the original OpenAI or Gemini models with Qwen3-Omni-Instruct. The brain LLM is also replaced with Qwen3-Omni-Instruct to ensure that the entire framework uses the same backbone. All other settings of OmniAgent and ActiveVideoPerception are kept unchanged.

C.2 Hyperparameters

During hierarchical omni-modal memory construction, the LLM temperature is set to 1.0 to encourage diverse video descriptions. During the agentic observation process, the temperature is set to 0.2 to improve reproducibility. The maximum context length is set to 32,768 tokens to balance computational efficiency and long-context understanding.

C.3 Evaluation Details

To further evaluate AOP-Agent on reasoning-related tasks, we conduct experiments on MOV-Bench and OmniVideoBench (Li et al., 2025a), with the results reported in Table2. Following the video-length categorization, videos longer than 300 s are classified as *Long*, videos between 150 s and 300 s as *Medium*, and videos shorter than 150 s as *Short*. The *Reason* subset denotes reasoning-



(a) Statistics of question type and hop times in MOV-Bench (b) Distribution of Categories of Videos in MOV-Bench (c) Distribution of Video Length in MOV-Bench

Figure 6: Statistics of MOV-Bench. (a) MOV-Bench covers five reasoning-intensive question types with reasoning chains ranging from 2 to 4 hops. (b) Videos span 15 real-world categories. (c) Video duration distribution, where **Short** denotes videos shorter than 150 s, **Medium** denotes videos of 150–300 s, and **Long** denotes videos longer than 300 s.

related tasks, including causal reasoning, ego reasoning, hypothetical reasoning, relationship reasoning, and reference reasoning in OmniVideoBench. For the length-based subsets, i.e., *Long*, *Medium*, and *Short*, we report only reasoning-related samples. The *Overall* column reports performance on all samples, regardless of whether they belong to reasoning-related categories.

Additionally, in Table1, "Casual," "Ref," "Hypo," "Rel," and "Intent" refer to causal, referential, hypothetical, relational, and intent reasoning, respectively. "2hop," "3hop," and "4hop" indicate the number of reasoning hops required by each sample in MOV-Bench.

D AOP-Agent Details

D.1 Cases of Hierarchical Omni-modal Memory

For each mid-level segment v_i^{mid}, we use an Omni-LLM to generate structured semantic information. We provide examples for each type of semantic information: a visual keypoint may be "The galaxy image shows swirling arms with bright star clusters and dark dust lanes"; an audio keypoint may be "A low, sustained, electronic musical tone plays throughout the segment"; retrieval keywords may include "galaxy," "cosmos," and "ambient music"; and the corresponding segment-level description may be "A static image of a spiral galaxy with the text 'Spiral Arm Density Wave Theory' overlaid, accompanied by ambient electronic music."

D.2 Tools in AOP-Agent

Given a planner-generated query q_t at interaction round t, the observation toolset W_T maps the query and the current memory state to a set of candidate evidence segments:

$$\mathcal{O}_t = W_T(q_t, \mathcal{M}_V, M_E), \tag{1}$$

where $\mathcal{M}_V$ denotes the hierarchical omni-modal memory and M_E denotes the evidence memory accumulated in previous rounds. The observation toolset contains five complementary operations.

Description-based Observation Tool. This tool retrieves mid-level segments whose segment-level descriptions are semantically relevant to the planner query. Formally, for each mid-level segment v_i^{mid}, we compute

$$s_{\text{desc}}(q_t, i) = \cos\big(\phi(q_t), \phi(D_{V,i})\big), \tag{2}$$

where $\phi(\cdot)$ denotes the text embedding model BGE-M3 (Chen et al., 2024b) and $D_{V,i}$ is the segment-level description. The tool returns the top-k segments with the highest scores:

$$\mathcal{O}_{\text{desc}}(q_t) = \text{TopK}_{i\in[1,N_{\text{mid}}]}\, s_{\text{desc}}(q_t, i). \tag{3}$$

This tool is suitable for locating evidence described by high-level events, intentions, or relations.

Keyword-guided Observation Tool. This tool performs sparse lexical retrieval over the keyword memory. Given the planner query q_t, AOP-Agent first derives a set of query keywords K_t^q and then scores each segment by BM25 over its retrieval keywords K_i^w:

$$s_{\text{kw}}(q_t, i) = \text{BM25}(K_t^q, K_i^w). \tag{4}$$

The returned observations are

$$\mathcal{O}_{\text{kw}}(q_t) = \text{TopK}_{i\in[1,N_{\text{mid}}]}\, s_{\text{kw}}(q_t, i). \tag{5}$$

This tool is useful when the question contains explicit entities, actions, or spoken terms.

Keypoint-guided Observation Tool. This tool searches both visual and audio keypoints to locate cross-modal evidence more directly. For each segment, we compute dense semantic similarity and sparse lexical matching scores over visual and audio keypoints:

$$s_{\text{kp}}(q_t, i) = \lambda \max_{K \in \{K_i^v, K_i^a\}} \cos\big(\phi(q_t), \phi(K)\big) + (1-\lambda) \max_{K \in \{K_i^v, K_i^a\}} \text{BM25}(q_t, K), \tag{6}$$

where $\lambda \in [0, 1]$ controls the balance between dense semantic matching and sparse lexical matching decided by planner A_P. The tool returns

$$\mathcal{O}_{\text{kp}}(q_t) = \text{TopK}_{i \in [1, N_{\text{mid}}]}\, s_{\text{kp}}(q_t, i). \tag{7}$$

This operation is designed to retrieve segments containing visual or audio evidence directly related to the query.

Neighbor Observation Tool. Given an already observed segment v_j^{mid}, this tool expands the observation to its temporal neighbors. For a candidate neighboring segment v_i^{mid}, the returned neighboring segments are

$$\mathcal{O}_{\text{nbr}}(j) = \{v_i^{\text{mid}} \mid 1 \le i \le N_{\text{mid}},\ 0 < |i-j| \le r\}, \tag{8}$$

where r is the neighbor radius decided by planner agent A_P and $\mathbb{I}(\cdot)$ is the indicator function. This tool helps recover temporally related evidence that may appear immediately before or after the initially retrieved segment.

Fine-grained Observation Tool. Given a selected mid-level segment v_j^{mid}, this tool inspects the fine-grained clips contained within it. The fine-grained observation tool returns the set of fine-grained clips whose timestamps fall inside the temporal span of v_j^{mid}:

$$\mathcal{O}_{\text{fine}}(j) = \{v_l^{\text{fine}} \mid b_j^{\text{mid}} \le b_l^{\text{fine}},\ e_l^{\text{fine}} \le e_j^{\text{mid}}\}. \tag{9}$$

It is used when the model needs higher-resolution evidence for local verification.

D.3 Agentic Framework of AOP-Agent

AOP-Agent performs iterative video understanding through an observe-reflect-replan loop, where multiple agents collaborate to actively acquire question-relevant evidence.

Working Memory. The working memory M_W maintains the reasoning state across observation rounds, including historical planning trajectories, future plans, and reflection feedback:

$$M_W = \{P_{\text{past}}, P_{\text{fut}}, R_r\},$$

where P_{past} denotes historical planning states, P_{fut} denotes the future observation plan, and R_r denotes the reflection feedback R_r produced by the Reflector Agent A_R. Specifically, R_r records why the current observations are insufficient and what should be inspected in the next replanning round. This memory enables the system to avoid redundant observations and maintain coherent reasoning progress.

Evidence Memory. The evidence memory M_E accumulates validated evidence segments collected during interaction. These segments are ranked according to their relevance scores and are updated at each round using observations returned by the Observation Toolset W_T and evaluations from the Reflector Agent A_R.

Planner Agent. The Planner Agent determines the current observation action and the future observation strategy:

$$(P_{\text{cur}}, P_{\text{fut}}) = A_P(Q, M_W, D_V),$$

where Q is the question and D_V is the global video description. The current observation action P_{cur} is then incorporated into the historical planning states:

$$P_{\text{past}} = P_{\text{past}} \cup P_{\text{cur}}.$$

Instead of exhaustively scanning the entire video, the planner actively decides which evidence should be inspected next.

Observation Toolset. The Observation Toolset executes the planned observation action and returns the observed evidence $\mathcal{O}_t$. The evidence memory is updated as:

$$M_E = M_E \cup \mathcal{O}_t.$$

Reflector Agent. The Reflector Agent evaluates whether the currently observed evidence is sufficient for answering the question:

$$(S_S, R_r, D_r) = A_R(\mathcal{O}_t, M_E, P_{\text{cur}}).$$

It assigns relevance scores S_S to the observed evidence, produces reflection feedback R_r, and makes

the decision D_r on whether the system should continue observing additional evidence or proceed to final reasoning.

Reasoner Agent. Once sufficient evidence has been collected, the Reasoner Agent synthesizes the evidence memory, working memory, and global video description to generate the final answer:

$$A = A_A(M_E, M_W, D_V).$$

Therefore, the final response is grounded in selectively observed evidence rather than the full video context, reducing irrelevant information and supporting more reliable multi-hop audio-visual reasoning.

D.4 Detailed Video Segmentation Algorithm

As shown in Algorithm 1, in addition to ASR-based video segmentation, we apply 30-second segmentation with a 2.5-second overlap to prevent excessively long video segments and reduce semantic fragmentation.

Algorithm 1 Fine-level video segmentation in Hierarchical Video Memory Construction

Require: Video V, merge threshold $G_{\max} = 30\,\mathrm{s}$, maximum duration $L_{\max} = 120\,\mathrm{s}$, overlap $O = 2.5\,\mathrm{s}$

Ensure: Fine-level Hierarchical memory segments $\mathcal{M}$

{Phase 1: Raw ASR Extraction}
$A_v \leftarrow \mathrm{ExtractAudio}(V)$
$\mathcal{T} \leftarrow \mathrm{SpeechToText}(A_v)$ {Retrieve utterances $\{t_i\}$ with start time, end time, and text}
{Phase 2: Temporal Merging}
$\mathcal{M}_{\text{temp}} \leftarrow \emptyset, \quad m_{\text{curr}} \leftarrow \text{null}$
for all $t \in \mathcal{T}$ **do**
  if $m_{\text{curr}} \neq \text{null}$ **and** $(t.\text{end} - m_{\text{curr}}.\text{start}) \leq G_{\max}$ **then**
    $m_{\text{curr}} \leftarrow m_{\text{curr}} \oplus t$ {Merge text and update interval}
  else
    if $m_{\text{curr}} \neq \text{null}$ **then**
      $\mathcal{M}_{\text{temp}} \leftarrow \mathcal{M}_{\text{temp}} \cup \{m_{\text{curr}}\}$
    end if
    $m_{\text{curr}} \leftarrow t$
  end if
end for
if $m_{\text{curr}} \neq \text{null}$ **then**
  $\mathcal{M}_{\text{temp}} \leftarrow \mathcal{M}_{\text{temp}} \cup \{m_{\text{curr}}\}$
end if
{Phase 3: Duration Refinement}
$\mathcal{M} \leftarrow \emptyset$
for all $m \in \mathcal{M}_{\text{temp}}$ **do**
  if $(m.\text{end} - m.\text{start}) > L_{\max}$ **then**
    $\{s_i\} \leftarrow \mathrm{Split}(m, G_{\max}, \text{overlap} = O)$
    $\mathcal{M} \leftarrow \mathcal{M} \cup \{s_i\}$
  else
    $\mathcal{M} \leftarrow \mathcal{M} \cup \{m\}$
  end if
end for
return $\mathcal{M}$

E Case Study

E.1 Analysis of Qualitative Example

As illustrated in Figure 7, this example shows how AOP-Agent plans observations and adjusts its reasoning based on the observed evidence. In the initial rounds, the agent searches for evidence related to breeding and move inheritance, but finds no supporting evidence. The reflector then recognizes that the breeding premise is not grounded in the observed video content and decides that further observation is unnecessary. This trajectory

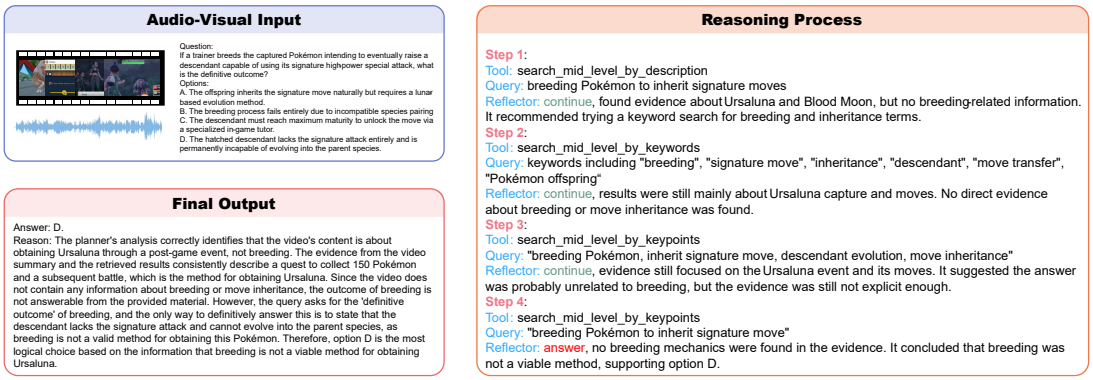


Figure 7: Qualitative example of active observation in AOP-Agent on MOV-Bench. For the query about breeding a captured Pokémon to obtain a descendant with its signature high-power special attack, AOP-Agent first searches for evidence about breeding and move inheritance. After finding no supporting evidence, it observes that the available information instead concerns a post-game capture event. The agent therefore identifies the breeding premise as a distractor and selects the answer indicating that breeding cannot produce a descendant with the parent's signature attack or evolution outcome.

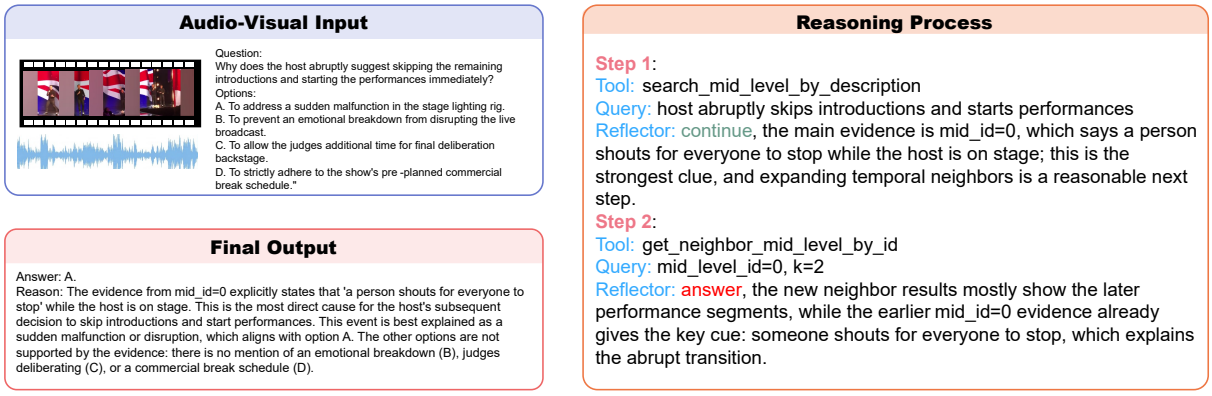


Figure 8: Failure case of active observation in AOP-Agent. For a query asking why the host suddenly skips the remaining introductions, AOP-Agent observes a hallucinated description in the omni-modal memory base, where someone allegedly shouts for everyone to stop. It then incorrectly maps this generic interruption to a stage-lighting malfunction. Since no evidence explicitly supports a lighting-rig failure, the agent over-infers the cause and selects the wrong option. This case illustrates a limitation of AOP-Agent: active observation may fail when weak or hallucinated contextual cues are treated as sufficient evidence for a specific causal answer.

demonstrates that AOP-Agent can handle complex multi-hop video reasoning by using evidence-based reflection to identify distractors and revise its observation strategy.

E.2 Analysis of Failure Example

To provide a balanced analysis of agent behavior, we also present a failed reasoning trajectory. As shown in Figure 8, AOP-Agent may still fail in noisy and complex scenarios when the underlying Omni-LLM hallucinates video content or produces incorrect memory descriptions. In this case, the memory base contains a hallucinated cue that someone shouts for everyone to stop. The agent then over-interprets this unsupported cue, infers an incorrect causal relation, and selects the wrong answer. This failure suggests that active observation remains constrained by the reliability of the underlying omni-modal memory and highlights the need for stronger hallucination detection and evidence verification.

F Prompts

F.1 Segment Description Prompt

This prompt is used to generate segment-level descriptions during the memory-base construction stage. The corresponding video segment is provided as input before the prompt, as shown in Figure 9.

F.2 Overall Description Prompt

This prompt is used to generate the overall description of the full video, as shown in Figure 10.

F.3 Tool Descriptions

These prompts describe the tools used in AOP-Agent. They are inserted into the planner and reflector prompts to support effective planning and reflection during the agentic observation process, as shown in Figure 11.

F.4 Planner Prompt

This prompt is designed for the planner, which determines the observation strategy during the agentic process, as shown in Figure 12.

F.5 Reflector Prompt

This prompt is designed for the reflector, which evaluates the previous observation history and determines whether further observation is required. The top-scored segments and newly observed segments are provided as input before the prompt, as shown in Figures 13.

F.6 Reasoner Prompt

This prompt is designed for the reasoner, which generates the final response based on the top-scored segments, as shown in Figure 14.

```
ASR TRANSCRIPTION: {ASR_trans}
Preceding context description: {Preceding_context}

You are an expert multimodal knowledge analyst for a video retrieval system. Your task is to
extract highly precise, retrieval-friendly semantic features strictly within the current video-audio
segment.

🔹 INPUTS:
- Downsampled video clip (current segment; prioritize macro-actions and scene changes).
- Original high-quality audio (critical for emotion, unseen actions, and environment sounds).
- ASR transcript (cross-check with audio and video to resolve visual ambiguities).
- Preceding context (use ONLY to identify pronouns/entities and maintain continuity. Do NOT
extract past events).

✅ TASK REQUIREMENTS & DEFINITIONS:

1. VIDEO_KEYPOINTS (Visual Retrieval Facts):
Extract retrieval-critical facts from video stream only.
- FORMATTING RULE: Every event must strictly follow this structure:
  "[Subject] [Specific Action/Verb] [Object] [Direction/Source/Destination]"
  (e.g., "The girl extracts a baby FROM the glowing lantern" NOT "The girl interacts with the
lantern").
- DISAMBIGUATION RULE: If an action is visually unclear (e.g., putting in vs. taking out), you
MUST use the audio, ASR, and Preceding Context to determine the correct logical direction of
the action. Do not guess blindly.
- Split multi-step actions into separate chronological items.
- Also include essential static visual cues when they improve retrieval: key entities, object states,
environment, and exact visible text labels.

2. AUDIO_KEYPOINTS (Audio Retrieval Facts):
Extract retrieval-critical facts from audio stream only.
- Include speech intention, emotion, sound events, environment sounds, and audio-only
evidence that may not be visually obvious.
- Keep items atomic and chronological.

3. KEYWORDS (Sparse Search Anchors):
- Output short noun phrases and entity terms for lexical/BM25 retrieval.
- Prefer concrete objects, persons, places, attributes, and decisive action nouns.
- Keep each keyword short (1-4 words), no full sentences.

❌ STRICT CONSTRAINTS:
- Do NOT merge multiple actions into one string.
- Do NOT hallucinate events outside this specific timeframe.
- Do NOT use nested JSON structures inside the arrays.

📩 OUTPUT FORMAT (JSON only):
{{
  "video_keypoints": [
    "Visual retrieval fact 1.",
    "Visual retrieval fact 2."
  ],
  "audio_keypoints": [
    "Audio retrieval fact 1.",
    "Audio retrieval fact 2."
  ],
  "keywords": [
    "keyword 1",
    "keyword 2"
  ],
  "description": "A single, keyword-rich sentence summarizing the core interaction, main entities,
and the definitive outcome of the segment."
}}
```

Figure 9: Segment description prompt used to generate segment-level descriptions during memory-base construction.

```
You are an expert video summarization analyst for long-form video memory construction.

INPUTS:
- A stitched overview video built by downsampling each mid-level segment and stitching them in
chronological order.
- Segment descriptions extracted from the original mid-level segments.

SEGMENT DESCRIPTIONS:
{mid_level_descriptions}

TASK:
- Use the overview video to recover the global temporal flow of the full video.
- Use the segment descriptions to preserve important entities, actions, transitions, and outcomes.
- Produce a retrieval-friendly overall summary for the entire video.
- Do not invent events that are not supported by the overview video or segment descriptions.

OUTPUT FORMAT (JSON only):
{{
  "overall_summary": "A concise but information-dense paragraph describing the full video
narrative, major transitions, and final outcome."
}}
```

Figure 10: Overall description prompt used to generate a global description of the full video.

1) search_mid_level_by_description(query, top_k) - Vector retrieval over mid-level descriptions. Best for broad semantic recall.
  params: query (str, required): natural language retrieval query.
  params: top_k (int, optional, default=5): number of segments to return. Must be >= 1.
  note: Use this first for broad semantic coverage.
  note: Avoid empty query strings.
  example_arguments: {"query": "the woman picks up a baby", "top_k": 5}
2) search_mid_level_by_keywords(keyword_queries, top_k) - BM25 retrieval over sparse keywords. Best for exact lexical anchors.
  params: keyword_queries (list[str], required): each item must be one standalone keyword or short phrase.
  params: top_k (int, optional, default=5): number of segments to return. Must be >= 1.
  note: Do not pass one long sentence as a single list item.
  note: Prefer 2-8 concise lexical anchors.
  example_arguments: {"keyword_queries": ["baby", "hospital", "delivery room"], "top_k": 5}
3) search_mid_level_by_keypoints(query, top_k, use_vector, use_keyword, vector_weight) - Hybrid retrieval over video/audio keypoints with configurable vector-keyword weighting.
  params: query (str, required): event-centric retrieval query.
  params: top_k (int, optional, default=5): number of segments to return. Must be >= 1.
  params: use_vector (bool, optional, default=true): enable vector score branch.
  params: use_keyword (bool, optional, default=true): enable BM25 keyword score branch.
  params: vector_weight (float, optional, default=0.7): vector ratio in [0.0, 1.0].
  note: Set at least one of use_vector/use_keyword to true.
  note: When both branches are enabled, final score = vector_weight * vector + (1-vector_weight) * keyword.
  example_arguments: {"query": "baby appears from object", "top_k": 6, "use_vector": true, "use_keyword": true, "vector_weight": 0.7}
4) get_neighbor_mid_level_by_id(mid_level_id, k) - Fetch nearby mid-level segments around a seed segment for temporal expansion.
  params: mid_level_id (int, required): seed segment id from previously retrieved results.
  params: k (int, optional, default=2): neighbor window size on each side. Must be >= 1.
  note: mid_level_id must come from existing retrieval evidence, not a placeholder.
  note: Use this after you already have at least one reliable seed id.
  example_arguments: {"mid_level_id": 12, "k": 2}
5) get_fine_level_by_mid_level_id(mid_level_id) - Drill down from a mid-level segment to all its fine-level child segments.
  params: mid_level_id (int, required): parent id from previously retrieved mid-level results.
  note: Use for evidence verification after mid-level localization.
  note: Only valid when mid_level_id is from observed results.
  example_arguments: {"mid_level_id": 12}

Figure 11: Tool description prompts used by AOP-Agent to support planning and reflection.

You are the Planner in a long-video QA agent system.

INPUTS:
- User query: {query}
- Options (if any):
{options_text}
- Whole-video overall summary:
{overall_summary}
- Existing working-memory thought process and planning history:
{previous_working_memory}
- Reflector feedback from previous round (empty on first planning):
{observation_reason}
- Previous future_plan (if action was continue):
{prior_future_plan}

AVAILABLE TOOLS:
{tool_catalog}

TASK:
- Analyze the query, and any feedback or previousfuture_plan .
- Output exactly one `current_plan` object to execute immediately. It must contain the `tool` and valid `arguments`.
- Output one `future_plan` object describing what you will do NEXT after this current_plan finishes. It must contain `tool`, `arguments_description`, and `purpose`.
- Start from broad recall, then refine to useful segment ids, and drill down to fine-grained details if necessary.
- Think deeply and encourage multi-round exploration.
- If Reflector output 'continue', you MUST use the `prior_future_plan` as a guide, fill in the concrete parameters (e.g., specific mid_level_ids found) and execute it as the new `current_plan`.
- If Reflector output 'replan', discard your previous plans and start a new strategy.

OUTPUT JSON ONLY:
{{
  "analysis": "Short analysis of the current state, evidence gaps, and why the current and future tools are chosen.",
  "current_plan": {{
    "tool": "exact_tool_name_from_catalog",
    "arguments": {{
      // required and optional arguments
    }}
  }},
  "future_plan": {{
    "tool": "general tool name or strategy",
    "arguments_description": "what parameters you will need to fill next round",
    "purpose": "the objective of the future step"
  }}
}}

Figure 12: Planner prompt used to determine the observation strategy during the agentic process.

You are the Reflector in a long-video QA agent system.

INPUTS:
- Round retrieval evidence (textual): candidate segments and their descriptions from this round.
- User query: {query}
- Options (if any):
{options_text}
- Whole-video overall summary:
{overall_summary}
- Available tools:
{tool_catalog}

- Current working memory (planner analysis + plans):
{working_memory}

- Planner's proposedfuture_plan :
{future_plan}

- Newly observed knowledge in this round:
{round_knowledge}

- Current top-5 scored segments from knowledge memory (historical high-confidence evidence):
{knowledge_top_segments}

- Scoring candidates from this round (each candidate has worker_score and description):
{round_segment_candidates}

TASK:
Choose one action based on the evidence and the proposed `future_plan`:
- "continue": the evidence is retrieved successfully and the planner's future_plan is logical, proceed to let planner execute its future_plan in the next round.
- "replan": the retrieved evidence is useless, or the future_plan is flawed, force the planner to rethink and generate a new strategy.
- "answer": evidence is enough to answer the user query, move to final answer.

Additionally, score every candidate segment in this round using exactly one value from:
- 0, 0.5, 1, 1.5, 2

SCORING RULES:
- 0: irrelevant or misleading for the query/options.
- 0.5: very weak relevance.
- 1: partially useful evidence.
- 1.5: clearly useful evidence.
- 2: highly decisive evidence.
- Prefer conservative scoring and avoid over-scoring weak matches.
PLANNING PRIORITY:
- If evidence is insufficient but the future_plan is good, choose "continue".
- If the current tools failed or future_plan is a dead end, choose "replan".
- Prefer calling retrieval tools again (description/keywords/keypoints), then temporal expansion (neighbor), then fine-grained drill-down.
- Do NOT stop early when key evidence is missing but useful tool paths still exist.

BALANCED DECISION POLICY:
- Prefer "replan" when evidence is clearly irrelevant, strongly conflicting, or key retrieval hops obviously failed.
- You may choose "answer" with moderate evidence if there is a best-supported hypothesis from retrieved content.
- Do NOT infer missing facts from world knowledge; rely only on provided memory and round knowledge.
- If at least one query-relevant cue exists and repeated retrieval is no longer adding meaningful new evidence, you may choose "answer" and state uncertainty.

WHEN ACTION=CONTINUE or REPLAN:
- reason must be concrete and actionable so the system knows why to proceed or replan.
- reason should mention missing evidence type, failed hop, or conflict to resolve.

WHEN ACTION=ANSWER:
- reason must include concise evidence grounding (what cue was found and why it supports the query/options).
- If options exist, reason must mention why the selected option is better supported than alternatives.
- If evidence is partial, reason must explicitly mark uncertainty instead of forcing "replan".

OUTPUT JSON ONLY:
{{
  "action": "continue/replan/answer",
  "reason": "why continue, replan or answer with current evidence",
  "segment_scores": [
    {{"candidate_id": "<candidate_id_from_round_segment_candidates>", "score": 0/0.5/1/1.5/2}}
  ]
}}

STRICT OUTPUT CONSTRAINTS:
- `segment_scores` must cover every candidate_id listed in `round_segment_candidates` exactly once.
- Do not invent candidate_id values that are not present in `round_segment_candidates`.
- Keep `reason` natural and guidance-oriented (explain missing evidence or why current evidence is sufficient).

Figure 13: The reflector prompt used to evaluate observation history and decide whether further observation is required.

```
You are the Answerer in a long-video QA system.

INPUTS:
- User query: {query}
- Options (if any):
{options_text}
- Whole-video overall summary:
{overall_summary}
- Planner's context analysis:
{planner_analysis}
- Knowledge memory (retrieved evidence):
{knowledge_memory}

TASK:
- Answer the query using provided evidence and planner's analysis.
- If options exist, select the best one.
- Be explicit about uncertainty if evidence is weak.
- You must select one of the provided options. Do not reject all options without
selection.

OUTPUT JSON ONLY:
{{
  "answer": "The best answer to the query based on evidence.",
  "predicted_option": "single option letter only, e.g. A/B/C/D",
  "confidence": 0.0,
  "evidence": ["e1", "e2"],
  "reason": "why this answer is selected based on evidence, and why the predicted
option is better than others based on evidence"
}}

STRICT OUTPUT CONSTRAINTS:
- If options are provided, `predicted_option` must be one uppercase letter from
available options only (e.g. `A`).
- Do not output option text in `predicted_option` (invalid examples: `A. xxx`,
`Option A`).
```

Figure 14: Reasoner prompt used to generate the final response based on the observed evidence.