OMNIAGENT: ACTIVE PERCEPTION AGENT FOR OMNIMODAL AUDIO-VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

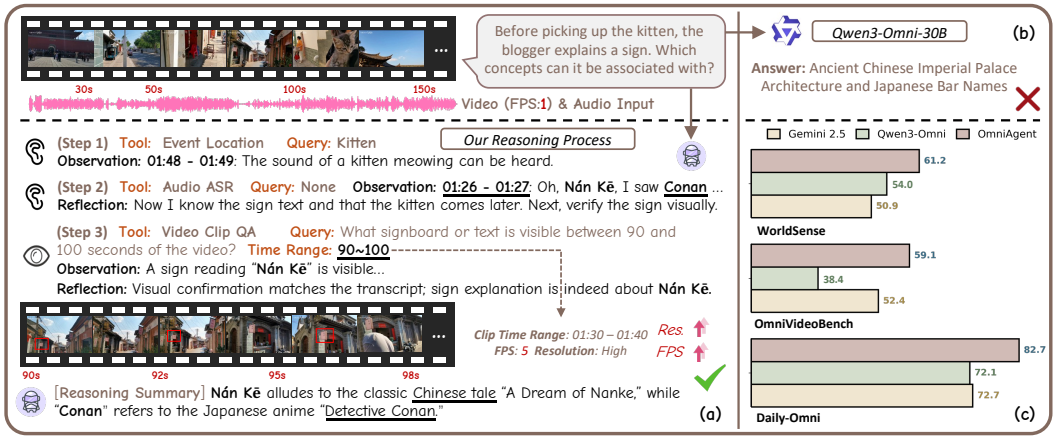

Figure 1: **(a)**: Illustration of OmniAgent, an active perception agent designed for omnimodal understanding. Given a user query, our agent employs a recursive *"Think-Act-Observe-Reflect"* loop and actively orchestrates multimodal tools (video, audio, and event tools) for fine-grained audio-video understanding. Due to the unavailability of the thinking process, we use reflection in the final summary of the agent to better understand. The presented video clip is from a vlog; the question is about two Chinese characters on a hanging signboard in the video. Initially, the agent utilizes audio to locate the temporal segment with the key information ("the kitten"), then invokes the video clip tool within that time window. Within the salient segment, we can afford model inference at an *increased* spatial and temporal resolution. With sufficient relevant visual evidence and the audio as input, the agent derives the correct answer. **(b)**: In contrast, the end-to-end model Qwen3-Omni (Xu et al., 2025b) cannot achieve such fine-grained reasoning and gives the wrong answer. **(c)**: Performance comparison on three audio-video understanding benchmarks. OmniAgent demonstrates superior performance without training, consistently outperforming strong end-to-end OmniLLMs such as Qwen3-Omni and Gemini 2.5-Flash (Comanici et al., 2025).

ABSTRACT

Omnimodal large language models have made significant strides in unifying audio and visual modalities; however, they often face challenges in fine-grained cross-modal understanding and have difficulty with multimodal alignment. To address these limitations, we introduce **OmniAgent**, to the best of our knowledge, the first fully active perception agent that dynamically orchestrates specialized unimodal tools to achieve more fine-grained omnimodal reasoning. Unlike previous works that rely on rigid, static workflows and dense frame-captioning, we demonstrate a paradigm shift from passive response generation to active multimodal inquiry. OmniAgent employs dynamic planning to autonomously orchestrate tool invocation on demand, strategically concentrating perceptual attention on task-relevant cues. Central to our approach is a novel coarse-to-fine audio-guided perception paradigm, which leverages audio cues to localize temporal events and guide subsequent reasoning. Extensive empirical evaluations on three audio-video understanding benchmarks demonstrate that OmniAgent achieves state-of-the-art performance, surpassing leading open-source and closed-source models by substantial margins of 10% - 20% accuracy without training.

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.

1 INTRODUCTION

Recently, end-to-end omnimodal large language models (OmniLLMs) have achieved encouraging results by integrating visual and audio encoders into a unified architecture (Tang et al., 2025; Zhang et al., 2024c; Xu et al., 2025a;b; Yang et al., 2025b; Ge et al., 2025; Shu et al., 2025; Yang et al., 2025c; AI et al., 2025; Team, 2026b;a). Despite this progress, as shown in Figure 2(a), OmniLLMs still face challenges in fine-grained cross-modal understanding, and the joint alignment training of audio and video representations poses significant challenges (Gallougah et al., 2025; Chowdhury et al., 2024; Sun et al., 2023a). Consequently, models often cannot respond accurately.

A critical empirical observation is that while MLLMs have demonstrated exceptional proficiency in unimodal tasks (Wang et al., 2025a; Team et al., 2025a; Bai et al., 2025b; Wang et al., 2024a; Chu et al., 2023b; Ding et al., 2025a; Bai et al., 2025a; Feng et al., 2025c;b), cross-modal understanding remains constrained by the challenges of temporal and feature alignment (Chowdhury et al., 2024; Jiang et al., 2025; Fan et al., 2025; Tao et al., 2026). Consequently, developing an agent to synergize the capabilities of distinct modalities is now a promising direction. Previous omnimodal agents rely predominantly on static workflows (Cao et al., 2025; Zhou et al., 2025), as illustrated in Figure 2(b). These methods face challenges in effectively harnessing the inherent reasoning capabilities of models for dynamic planning, thereby impeding the attainment of a fine-grained understanding.

Recent works have yielded significant advancements in agent-based video-only understanding (Zhang et al., 2025b; 2024b; Fan et al., 2024; Wang et al., 2024b; Yang et al., 2025d; Wang et al., 2025d; Pang & Wang, 2025; Wang et al., 2025e; Chowdhury et al., 2025; Yin et al., 2025; Tian et al., 2025). Specifically, temporal event localization is paramount for fine-grained analysis. Prevailing approaches predominantly rely on frame-captioning, where captions are generated for sampled frames, stored, and subsequently retrieved and analyzed iteratively by agents (Wang et al., 2024b; Zhang et al., 2025b; Wang et al., 2025d). While these methods refine their hypotheses through multi-step inference, they incur substantial computational overhead. Moreover, the generated captions may occasionally be irrelevant to the query. However, in the context of audio-visual understanding, the audio modality presents distinct challenges yet offers a unique opportunity: unlike redundant visual signals, audio naturally provides accurate and concise temporal grounding information regarding the salient events (Guo et al., 2025; Tao et al., 2025b; Wu et al., 2025; Chen et al., 2025; Xie et al., 2025; Lin et al., 2025). This information is efficiently utilized for event localization and to emulate a reasoning process akin to human cognition, thereby facilitating more comprehensive reasoning.

Building on these insights, we present OmniAgent, an agent specifically designed for omnimodal (audio-visual) understanding in an active reasoning fashion, which treats strong single-modal models as callable tools. In contrast to previous methods constrained by fixed workflows, as shown in Figure 2(c), our approach initiates a fundamental paradigm shift from *passive* response generation to *active* information inquiry. The agent employs an LLM as the central component to orchestrate tool invocation, determining the optimal modality to use – it explicitly decides *whether* to attend to audio or video, and *how* to process the information. Note that this process is completely *autonomous*, decided by the LLM itself. The tool calling and decision-making are transparent to us, thus explainable and optimizable. By strategically selecting spatial and temporal focus range (*i.e.*, deciding precisely where to look and listen), OmniAgent achieves genuine, fine-grained understanding.

Specifically, we construct a comprehensive library of tools categorized into three distinct sets: (1) Video tools, (2) Audio tools, and (3) Event tools. The *video toolset* enables global captioning and general visual QA, while also allowing for the analysis of specific temporal windows at higher sampling rates to support more fine-grained understanding. The *audio toolset* incorporates audio captioning and detailed QA capabilities, complemented by timestamped automatic speech recognition (ASR) for precise speech grounding. Within the *event toolset*, we propose a novel *audio-based event localization* strategy. This mechanism empowers the agent to autonomously query and temporally localize events across the entire audio stream, establishing temporal anchors for subsequent fine-grained analysis. By synthesizing the capabilities of distinct MLLMs, our agent adaptively leverages their complementary strengths to facilitate joint, fine-grained analysis, utilizing cross-modal corroboration to maximize audio-visual comprehension performance.

Extensive experiments show that OmniAgent achieves the *best* accuracy on several audio-video understanding benchmarks, surpassing the state-of-the-art open-source and closed-source models,

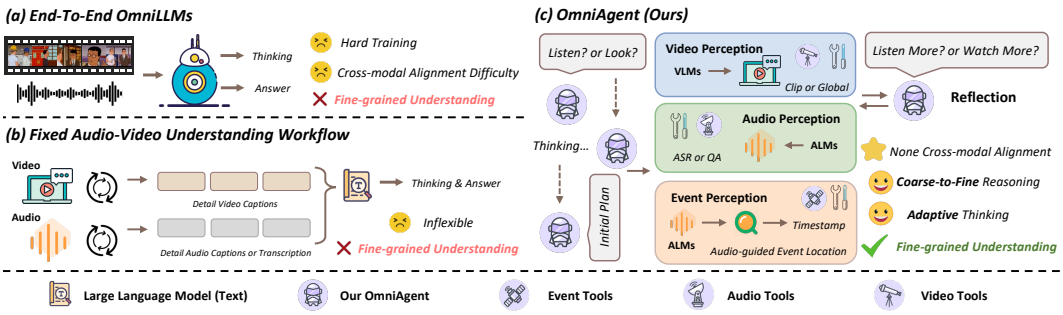

Figure 2: *(a) End-to-end OmniLLMs* implicitly fuse modalities but suffer from high training costs, difficult alignment, and limited fine-grained reasoning. *(b) Fixed workflow agents* rely on rigid pipelines, lacking the flexibility to allocate attention for fine-grained analysis adaptively. *(c) Our OmniAgent* employs active perception reasoning and inquiry. Within an iterative, reflective loop, the agent strategically leverages the ability to understand video and audio. This explicitly solves the cross-modal alignment difficulty and achieves fine-grained understanding.

such as Qwen3-Omni-30B (Xu et al., 2025b) and Gemini2.5-Flash (Comanici et al., 2025), by a significant margin of 10-20% accuracy without training.

The main contributions of this work are:

- We introduce *OmniAgent*, a novel agent-based framework tailored for comprehensive audio-video understanding. Employing an active perception strategy, it dynamically modulates attention between auditory and visual modalities and, via a self-reflective mechanism, solves the cross-modal alignment problem.
- We construct a comprehensive modality-specific toolkit and introduce an audio-guided event localization algorithm designed to facilitate *fine-grained cross-modal reasoning*.
- Experimental results on several audio-video understanding benchmarks show that OmniAgent achieves the new SoTA, with significant accuracy improvement compared with open-source and closed-source models.

2 Related Work

2.1 Omnimodal Large Language Models

End-to-end OmniLLMs aim to achieve a common understanding across all modalities, including image, audio, video, and text. By leveraging multimodal data, these architectures acquire richer contextual representations and gain a deeper understanding of inter-modal relationships (Xu et al., 2025a; Tong et al., 2025; Xu et al., 2025b; Xie & Wu, 2024; Tang et al., 2025; Zhang et al., 2024c; Yang et al., 2025b; Ge et al., 2025; Shu et al., 2025; Sun et al., 2024; Li et al., 2024; Team et al., 2025b). Recent works, such as Qwen3-Omni (Xu et al., 2025b) and the Video-SALMONN series (Sun et al., 2024; Tang et al., 2025), have introduced state-of-the-art end-to-end models capable of unified multimodal perception. Among closed-source models, Gemini (Comanici et al., 2025; Team et al., 2024; 2023) stands as a powerful baseline, distinguished by its strong multimodal understanding capabilities. However, end-to-end models face significant hurdles in achieving fine-grained cross-modal understanding.

2.2 Video Understanding Agents

Leveraging the advanced capabilities of MLLMs, recent studies have investigated agentic approaches to address the intricacies of video understanding with video clip captioning (Wang et al., 2024b; Zhang et al., 2024a; Jeoung et al., 2024; Wang et al., 2025c; Park et al., 2024; Ma et al., 2025; Kahatapitiya et al., 2025; Jeoung et al., 2024; Kugo et al., 2025). Concurrently, other methodologies have focused on decomposing complex queries into multi-step processes utilizing specialized tool modules (Fan et al., 2024; Liu et al., 2025a; Min et al., 2024; Zhu et al., 2025b; Zhang et al., 2024b). More recently,

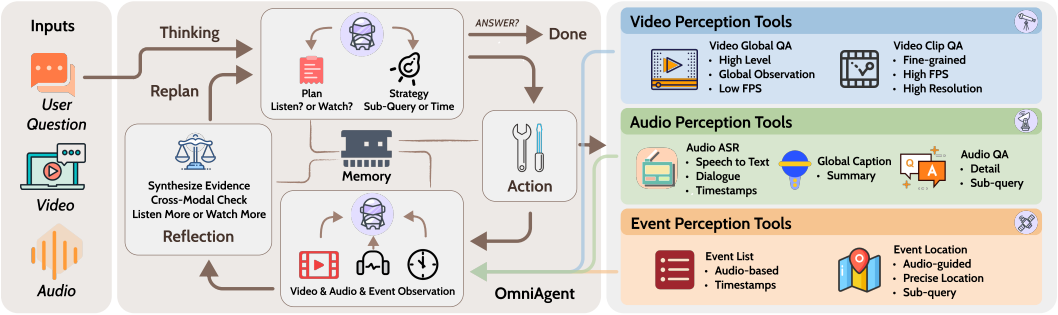

Figure 3: **Overview of the OmniAgent framework.** The system processes audio-video inputs through an iterative *thinking-action-observe-reflection* cycle. The agent utilizes a comprehensive suite of perception tools (video, audio, and event) to gather fine-grained evidence, while the reflection module synthesizes observations to update the memory and decide whether to rethink or answer.

research has shifted away from static workflows to explore active agentic perception (Yao et al., 2022; Yuan et al., 2025; Zhang et al., 2025b; Wang et al., 2025e; Gao et al., 2025; Yang et al., 2025a), thereby enhancing long video comprehension. However, comprehensive omnimodal understanding remains challenging due to the complexities of cross-modal alignment and fine-grained reasoning.

3 OmniAgent

We introduce the *OmniAgent*, to the best of our knowledge, the *first* active perception agent framework for omnimodal understanding. In contrast to conventional paradigms that rely on passive frame processing or rigid execution protocols, OmniAgent dynamically orchestrates a suite of modality-specific perception tools, effectively reformulating audio-video understanding from a passive retrieval task into an active, sequential decision-making process.

3.1 Overview and Problem Formulation

Motivation. Existing end-to-end OmniLLMs typically process video and audio streams by projecting them into a shared latent space. However, this paradigm exhibits a fundamental limitation: the inability to allocate attentional resources between modalities adaptively. Crucially, query-relevant information is often modality-specific; valid responses may hinge exclusively on auditory cues or demand scrutiny of high-resolution visual details. Constrained by fixed token budgets and joint optimization objectives, OmniLLMs lack the architectural flexibility to dynamically prioritize specific modalities or adjust processing granularity. This deficiency frequently results in the degradation of fine-grained understanding.

Formulation. To address this, we formulate omnimodal understanding not as a static task, but as a sequential, active decision-making process. Let $\mathcal{V}$ and $\mathcal{A}$ denote the visual and audio streams, and q be the user query. We define an agent π and store a memory $\mathcal{M} = \{a_0, o_0, \ldots, a_T, o_T\}$. At each step t, the agent assesses its state s_t and actively selects an action $a_t \in \mathcal{T}$ and gets an observation o_t (e.g., *Listen* to a segment or *Watch* a specific region), to maximize the information gain regarding q. By explicitly decoupling the modalities into callable tools, OmniAgent empowers the model to autonomously determine the optimal modality and granularity—deciding when to rely on low-cost auditory cues and when to demand high-cost visual inspection—thereby solving the cross-modality alignment difficulty.

3.2 Modality-Aware Expert Toolset

To facilitate precise interaction with the environment, we devise a comprehensive toolset, denoted as $\mathcal{T}$, stratified by both modality and granularity. Functioning as the perceptual interfaces of the agent, these tools offer varying degrees of information density and computational overhead.

Video Perception Tools ($\mathcal{T}_V$). While visual processing yields rich semantic information, it incurs high computational costs. Relying solely on global representations often leads to the loss of

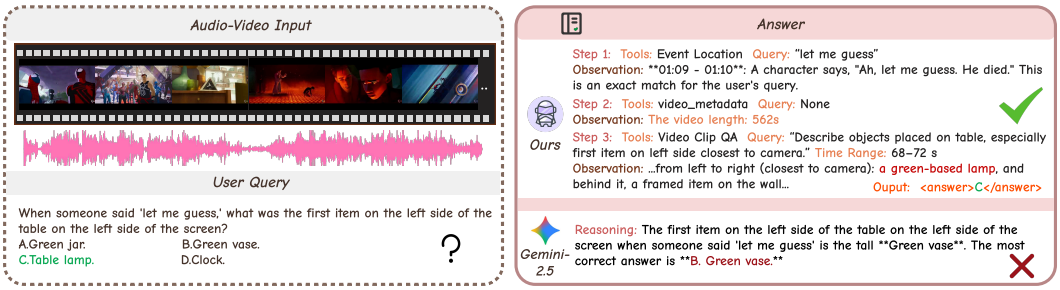

Figure 4: Visualization of the responses and underlying reasoning processes generated by our OmniAgent and Gemini2.5-Flash to an audio-video understanding question.

fine-grained details. To address these trade-offs, we design two distinct visual tools: **`Global_QA`**: $\mathcal{T}_{\text{VGA}}$ and **`Clip_QA`**: $\mathcal{T}_{\text{VCA}}$. For $\mathcal{T}_{\text{VGA}}$, we employ sparse frame sampling to mitigate the overhead of long sequences. Additionally, this tool allows the agent to identify initial visual cues for coarse temporal localization. Conversely, $\mathcal{T}_{\text{VCA}}$ serves as the fine-grained analysis engine. It extracts video slices within a target temporal window and employs a higher sampling rate and input resolution. This enables deep visual reasoning, facilitating the detailed analysis of object actions and spatial relationships, while maintaining a balanced computational budget.

Audio Perception Tools ($\mathcal{T}_{\text{A}}$). Audio signals provide dense, complementary information essential for holistic video reasoning. First, the **`ASR`**: $\mathcal{T}_{\text{ASR}}$ transcribes spoken dialogue into text with precise timestamp alignment. This capability is indispensable for queries dependent on specific verbal cues or semantic narratives conveyed through speech. Second, the **`Global_Caption`**: $\mathcal{T}_{\text{AGC}}$ synthesizes a summary of the acoustic environment, establishing a global auditory context. Finally, the **`Audio_QA`**: $\mathcal{T}_{\text{AQ}}$ tool empowers the agent to formulate targeted inquiries, extracting specific acoustic details required for understanding.

Event Perception Tools ($\mathcal{T}_{\text{E}}$). Video understanding faces significant hurdles due to the hardness of event grounding over long sequences, rendering precise event localization a persistent challenge. We identify the audio modality as a pivotal opportunity to address this bottleneck; unlike video, audio captures global context and event semantics with low cost. Leveraging this efficiency, we propose an audio-guided event localization method. Specifically, the **`Event List`**: $\mathcal{T}_{\text{EL}}$ tool processes the entire audio stream to extract a discrete list of detectable sound events, enabling the agent to discern the global semantic context. Complementarily, **`Event Location`**: $\mathcal{T}_{\text{ELO}}$ accepts specific queries to return precise timestamps. This serves as an effective temporal proposal mechanism, allowing the agent to pinpoint occurrence times efficiently.

3.3 Agentic Design

To fully exploit the intrinsic reasoning and planning capabilities of the LLMs, we eschew rigid workflows or prescriptive tool usage. Instead, we formulate an iterative *Think-Act-Observe-Reflect* cycle, empowering the agent to actively orchestrate reasoning, planning, and execution across both modalities (*audio and video*), as shown in Figure 3.

Active Thinking. Upon receiving an input query, our agent formulates a strategic inference plan designed to maximize accuracy while utilizing information from both modes. Crucially, the system assesses the cross-modal dependency of the query to prioritize the optimal modality dynamically—determining whether to employ a "listen" or "watch" strategy—and selects the appropriate retrieval tools accordingly. At step t, we have:

$$a_t, \text{args}_t = \pi_{\text{plan}}(q, \mathcal{M}_t), \tag{1}$$

where the agent maintains a comprehensive contextual memory $\mathcal{M}_t$, aggregating the initial query, sequential observations, and accumulated background information to facilitate robust multi-step reasoning. args_t is the corresponding action parameter of a_t, such as the sub-question for $\mathcal{T}_{\text{AQ}}$ or $\mathcal{T}_{\text{VGA}}$ and the video clip start time & end time for $\mathcal{T}_{\text{VCA}}$.

Action & Observation The selected tool by the agent is executed on the corresponding modality streams:

$$o_t = \text{Execute}(a_t, \text{args}_t, \mathcal{V}, \mathcal{A}), \tag{2}$$

Table 1: **Comparison of different models on the Daily-Omni benchmark**. The **best** result for each metric is in bold, and the second-best is underlined. 'A' denotes the incorporation of audio modalities, whereas 'V' indicates reliance on visual modalities extracted from video inputs. The symbol '$*$' signifies that the model employs Chain-of-Thought (CoT) reasoning or extended inference before generating an answer.

Method	Modality	AV Event Alignment	Comparative	Context Understanding	Event Sequence	Inference	Reasoning	30s Subset	60s Subset	Avg
Closed-source Models										
GPT-4o	V	47.90	62.60	52.33	52.61	66.23	66.29	55.64	57.45	56.47
Qwen3-VL-Plus	V	51.68	77.10	61.66	66.99	71.43	68.57	63.68	66.55	65.00
Gemini 2.0-Flash	A+V	62.18	73.28	63.73	63.72	76.62	75.43	67.23	68.55	67.84
Gemini 2.5-Flash*	A+V	-	-	-	-	-	-	-	-	72.70
Open-source Models										
Unified-IO-2 XXL-8B	A+V	25.63	31.30	26.42	25.82	35.06	29.71	26.74	30.00	28.24
VideoLLaMA2-7B	A+V	35.71	35.88	35.75	31.70	40.91	34.29	38.02	31.82	35.17
Qwen2.5-Omni-7B	A+V	44.12	51.15	38.86	40.52	57.79	61.71	46.68	48.36	47.45
Ola-7B	A+V	40.34	61.07	40.41	43.46	63.64	69.71	51.47	49.82	50.71
Qwen3-Omni-30B	A+V	61.90	79.25	69.47	65.32	82.67	85.92	71.28	74.29	72.08
Agent-based Methods										
DVD	V	49.32	57.47	57.12	58.45	70.37	63.24	56.31	62.41	59.22
Daily-Omni	A+V	51.68	68.70	60.10	53.92	78.57	71.43	63.99	59.27	61.82
XGC-AVis	A+V	63.50	77.10	68.40	64.40	**85.10**	82.30	71.60	71.50	71.50
OmniAgent (Ours)	A+V	**80.67**	**83.21**	**80.83**	**81.05**	83.36	**86.86**	**80.37**	**85.45**	**82.71**

where o_t is the output of tools with a text response or a timestamp. The model will update the perception of the entire audio and video based on the initial thought derived from the observation of both modalities. And the memory is updated: $\mathcal{M}_{t+1} = \mathcal{M}_t \cup \{(a_t, o_t)\}$.

Reflection & Rethinking. Before the next iteration, the agent critically assesses all the acquired evidence. It determines the efficacy of the executed tool by synthesizing current outputs with historical context, using this data to refine the execution plan dynamically. Crucially, the module executes a cross-modal consistency check to identify potential discrepancies between visual and auditory signals. Consequently, if the agent determines that the accumulated multimodal evidence is insufficient to resolve the query, or if a cross-modal discrepancy arises necessitating further exploration, it reinitiates the thinking cycle. This iterative loop persists until the **`ANSWER`** operation is explicitly invoked, at which point the system synthesizes a final response and summary addressing the original user query.

Notably, *while OmniAgent predominantly leverages audio for event localization to achieve better time alignment, our framework can execute precise visual event localization when necessary and achieve better performance in unimodal-based reasoning* (see Figures 11 and 12 in the Appendix D.2 for more case analyses.). Figure 4 illustrates the reasoning process of our agent. In summary, we equip the agent with an evidence-based, reflective, and flexible action execution mechanism. Mirroring human cognitive processes, the system selectively extracts multimodal information and reasons about modal perception sub-problems. Consequently, this approach circumvents the computational complexity associated with rigid, dense cross-modal alignment. Through this iterative reasoning process, OmniAgent achieves fine-grained cross-modal understanding by synthesizing perceptions from both modalities, ultimately delivering superior accuracy in response to the given question.

4 Experimental Results

4.1 Experimental Settings

Benchmarks. We evaluate our method on three widely-used audio-video understanding benchmarks: Daily-Omni (Zhou et al., 2025), OmniVideoBench (Li et al., 2025a), and WorldSense (Hong et al., 2025). Daily-Omni primarily evaluates performance on short-form video segments of 30s and 60s, whereas OmniVideoBench comprehensively assesses audio-visual understanding capabilities in long-form videos. Complementarily, WorldSense gauges multimodal comprehension across eight distinct domains, focusing specifically on medium-length videos.

Compared Methods. We compare our agent with open-source MLLMs: VideoLLaMA2 (Cheng et al., 2024), Ola (Liu et al., 2025b), Unified-IO-2 (Lu et al., 2024), Qwen2.5-Omni (Xu et al., 2025a), Qwen2.5-VL (Bai et al., 2025b), Baichuan-Omni-1.5 (Li et al., 2025b), video-SALMONN 2 (Sun et al., 2024; Tang et al., 2025), Qwen3-VL (Bai et al., 2025a) and the state-of-the-art model

Qwen3-Omni (Xu et al., 2025b). And various closed-source MLLMs: Gemini2.5-Flash (Comanici et al., 2025), GPT-4o (OpenAI, 2023), Gemini2.0-Flash, and OpenAI o3. In addition, we also compare with the SoTA agent-based audio-video understanding framework: Daily-Omni (Zhou et al., 2025) and Xgc-avis (Cao et al., 2025). These approaches are predicated on static or semi-rigid agent workflows to varying degrees, standing in distinct contrast to the dynamic adaptability of our proposed framework. We also compare with the video understanding agent DVD (Zhang et al., 2025b).

Implementation Details. For the core of the agent, we use OpenAI o3 as the brain because of its excellent reasoning capabilities. We restrict the maximum number of iteration steps to 30. Regarding the component modules, we employ Qwen3-VL as the backbone for video perception and utilize Qwen3-Omni for audio global caption and ASR. Additionally, we select Gemini-2.5-Flash for the event perception tool and $\mathcal{T}_{\text{AQ}}$, capitalizing on its better time grounding capabilities. For more information, see Appendix A.

4.2 Comparison with SoTA Models

Table 1 presents the results on the Daily-Omni Benchmark. Specifically, OmniAgent substantially outperforms both proprietary baselines, such as Gemini-2.5-Flash-Thinking (72.7%), and SoTA open-source models like Qwen3-Omni (72.08%), achieving a remarkable overall accuracy of *82.71%*. This result validates that our agentic framework, by effectively synergizing specialized uni-modal capabilities, circumvents the inherent challenges of rigid cross-modal alignment. Furthermore, it demonstrates the efficacy of audio guidance in enhancing fine-grained audio-visual understanding. Relative to competing agent-based architectures, OmniAgent yields performance gains of *10%-20%*, underscoring the critical value of its active perception mechanisms.

Table 2: **Quantitative comparison on the OmniVideoBench**. The **best** result for each metric is in bold, and the second-best is underlined.

Method	Modality	(0,1] min	(1,5] min	(5,10] min	(10,30] min	Avg.
Closed-source Models						
Qwen3-VL-Plus	V	36.92	45.27	37.87	30.65	38.93
Gemini-2.0-Flash	A+V	49.40	43.15	41.05	34.87	41.50
Gemini-2.5-Flash*	A+V	55.42	55.10	47.37	52.11	52.40
Open-source Models						
Qwen2.5-VL-72B	V	33.13	30.03	31.88	24.43	29.50
VideoLLaMA2-7B	A+V	32.00	28.20	29.60	28.29	29.20
Qwen2.5-Omni-7B	A+V	41.57	27.41	25.33	26.72	29.30
Baichuan-Omni-1.5	A+V	28.92	31.78	28.38	32.44	30.70
Qwen3-Omni-30B	A+V	45.78	37.03	38.86	35.11	38.40
Agent-based Methods						
OmniAgent (Ours)	A+V	**66.08**	**58.53**	**59.03**	**55.64**	**59.10**

For long-video evaluation and more difficult questions, Table 2 presents results on the OmniVideoBench. The enhancement of OmniAgent over Qwen3-Omni-30B is notable, achieving an overall accuracy of *59.1%*. This performance substantially surpasses that of other open-source and closed-source end-to-end models, thereby further validating the efficacy of our agent algorithm. Figure 4 illustrates the comparative inference capabilities of OmniAgent against Gemini2.5-Flash, demonstrating that OmniAgent effectively resolves complex queries by leveraging active cross-modal reasoning. For medium-length video in the WorldSense benchmark, as shown in Table 3, it can also reflect the leading position of our OmniAgent. In the Appendix D, we provide more case studies about agent reasoning in different inputs.

4.3 Analyses on Reasoning Behaviors

The core LLM serves as the central reasoning engine within our OmniAgent. It autonomously synthesizes multimodal information to dynamically orchestrate tool execution. To elucidate these mechanisms, we conducted a quantitative analysis of tool invocation patterns across various LLMs on the Daily-Omni Benchmark (Zhou et al., 2025). Specifically, we measured both the distribution of tool calls and the associated average reasoning steps, as illustrated in Figure 5. In the Appendix Section B, we provide more behavior analyses and show the behavior of the agent in different benchmarks.

Finding 1. Across all LLM backbones, we observe a consistent strategic pattern: agents prioritize $\mathcal{T}_{\text{AGC}}$ (AGC) in the *initial phase* to establish a global contextual background. Subsequently, the reasoning process culminates with $\mathcal{T}_{\text{VCA}}$ (VCA), employed to extract the precise, fine-grained evidence necessary for the *final response*. This distinct sequential progression—from global audio context to

Table 3: **Comparison of different models on the WorldSense**. We compare our agent-based ( ) method against various baselines, including closed-source ( ) and open-source ( ) models. The **best** result for each metric is in bold.

Method	#Params	Modality	Tech & Science	Culture & Politics	Daily Life	Film & TV	Performance	Games	Sports	Music	Avg.
GPT-4o	-	V	48.0	44.0	38.3	43.5	41.9	41.2	42.6	42.7	42.6
Gemini 1.5 Pro	-	A+V	53.7	47.2	50.3	50.4	52.4	46.8	40.2	42.0	48.0
Gemini 2.5 Flash	-	A+V	55.1	48.2	53.0	48.8	56.2	47.2	46.3	50.0	50.9
Qwen2.5-Omni	7B	A+V	47.8	49.8	43.6	43.8	48.3	39.1	43.5	47.3	45.4
video-SALMONN 2+	72B	A+V	59.0	63.1	54.0	59.9	58.1	54.1	51.9	54.4	56.5
Qwen3-Omni	30B	A+V	-	-	-	-	-	-	-	-	54.0
OmniAgent	-	A+V	**64.3**	**66.3**	**59.4**	**63.1**	**62.2**	**59.2**	**55.8**	**60.3**	**61.2**

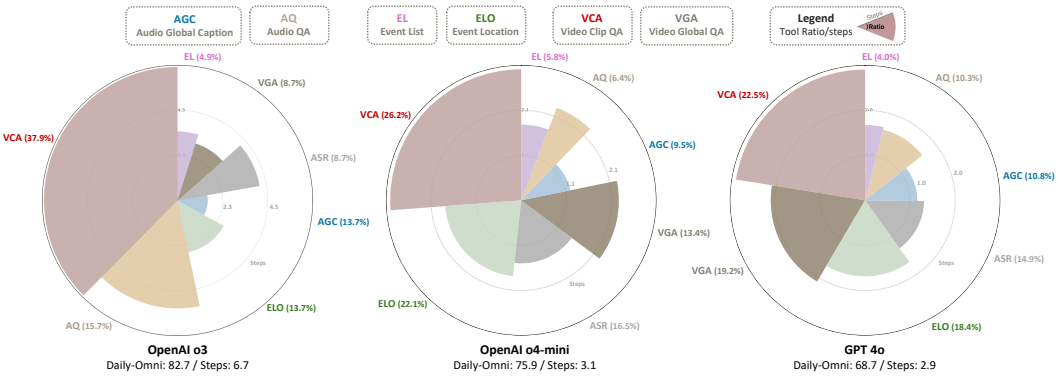

Figure 5: **Analysis of the behavior with different core LLM models.** We quantified tool utilization patterns by calculating both the proportion of invocations (call ratio) and the average number of reasoning steps per call. In the resulting visualization, the sector angle represents the *tool call ratio*, and the magnitude of the radius denotes the specific execution steps at which the tool was invoked.

localized visual verification—empirically validates the efficacy of our algorithmic design and the utility of the provided toolset.

Finding 2. The behavior of using OpenAI o3 as the LLM in our agent design aligns precisely with the core objective of cross-modal fine-grained understanding. We observe that the model preferentially utilizes granular tools—specifically $\mathcal{T}_{\text{VCA}}$ and $\mathcal{T}_{\text{AQ}}$—during the final resolution phase, while strategically deploying $\mathcal{T}_{\text{ELO}}$ for event localization during the intermediate reasoning stages. Conversely, computationally intensive tools ($\mathcal{T}_{\text{ELO}}$ and $\mathcal{T}_{\text{VGA}}$) provide macro-level insights, as they are unable to provide granular, fine-grained details. This progression effectively exemplifies the coarse-to-fine cognitive flow orchestrated by our agentic algorithm.

Finding 3. OpenAI o4-mini and GPT-4o exhibit a propensity for rapid convergence, frequently bypassing deeper reflection and iterative inspection phases. This tendency is particularly pronounced in GPT-4o, which demonstrates an excessive reliance on coarse-grained $\mathcal{T}_{\text{VGA}}$ outputs. Consequently, it fails to interrogate fine-grained visual details, resulting in low accuracy. Similarly, o4-mini displays a significant modality bias, disproportionately prioritizing visual information while neglecting the exploration of audio.

Insight. GPT-4o yields suboptimal performance, largely attributable to premature convergence on coarse-grained evidence. In contrast, o3 demonstrates a more deliberative process, effectively leveraging both modalities to unearth fine-grained details. This disparity underscores the critical necessity of the thinking-to-reflection cycle. Furthermore, our findings suggest that agentic systems must actively mitigate modal biases and strive for cross-modal consensus. Design protocols should prevent dominant visual signals from overshadowing critical auditory cues, thereby ensuring a holistic and accurate understanding of the video.

4.4 Ablation Study

Tool Model Choices. Table 4 (a) delineates the impact of backbone model selection for multimodal tools on the overall performance. Notably, the efficacy of the event model proves pivotal for the

Table 4: Ablation of tool model choices and multimodal tools. The results demonstrate the necessity of each tool in our agent.

Tools			Daily-Omni
Video Tool	*Audio Tool*	*Event Tool*	(Avg.)
Qwen3-VL	Qwen3-Omni	Gemini-2.5	82.7
Qwen3-VL	Qwen3-Omni	Gemini-2.0	74.2
Qwen2.5-VL	Qwen3-Omni	Qwen3-Omni	71.7
Qwen3-VL	Qwen2.5-Omni	Gemini-2.5	77.1
Gemini-2.5	Gemini-2.5	Gemini-2.5	**83.3**

(a) Tool Model Choices.

Multimodal Understanding Tools			Daily-Omni
Video Clip QA	*Audio QA*	*Event Location*	(Avg.)
	✓	✓	76.3
✓		✓	80.2
✓			77.3
✓	✓	✓	**82.7**

(b) Multimodal Tools.

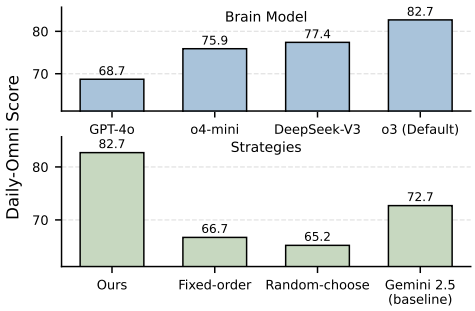

Figure 6: Ablation on different core LLM models and the framework effectiveness. We evaluate our agent across different brain models and compare the effects of fixed-order and random calls under the same tool setting, demonstrating the effectiveness of our overall framework design.

agent's reasoning capabilities. Given OmniAgent's substantial reliance on audio temporal grounding, Gemini-2.5 demonstrates superior proficiency in this domain. Conversely, previous Gemini iterations and open-source alternatives exhibit suboptimal temporal grounding. *This reminds us that focusing on and enhancing the audio-visual temporal grounding capabilities of OmniLLMs is a promising direction for future research* (discussed in detail in Appendix Section F). Furthermore, Qwen2.5-Omni (Xu et al., 2025a) suffers from performance degradation due to its inherent limitations in ASR and general audio comprehension. Surprisingly, employing Gemini 2.5-Flash across all tools yields better reasoning accuracy.

Multimodal Tools. Table 4 (b) details the ablation experiments conducted to evaluate the individual contributions of various tools. First, Video Clip QA proves indispensable for maintaining system accuracy. In its absence, the agent reverts to repetitive reliance on Global QA, consequently failing to resolve fine-grained details. Furthermore, the criticality of event tools is substantiated by the data. The exclusion of Audio QA and event localization tools precipitates a substantial performance decline, thereby validating the efficacy of the proposed agentic framework and tool design.

Core Models and Framework Design. Figure 6 illustrates the ablation experiments conducted on various core LLM models, detailing the inference accuracy achieved when utilizing an identical tool model and a constant call count. The results demonstrate that SoTA accuracy can also be attained by employing DeepSeek-V3 (Liu et al., 2024), an open-source LLM, thereby verifying the effectiveness of the design of our framework. Subsequently, when keeping the model constant, replacing our agent framework with fixed or random scheduling mechanisms results in a substantial degradation in accuracy, further highlighting the contribution of the design of the agent framework. Furthermore, we emphasize that the agent framework necessitates foundational capabilities from both the core model and the tool model; a core model lacking adequate reasoning abilities and a tool model exhibiting weak perception cannot successfully execute the processes designed within the agent framework.

5 Conclusion

We introduce OmniAgent, a fully active perception agent tailored for omnimodal audio-visual reasoning. Operating via a recursive *"Think-Act-Observe-Reflect"* loop, the system actively orchestrates tools to accumulate multimodal evidence, facilitating fine-grained comprehension progressively. Departing from conventional static workflows, we integrate a novel audio-driven event localization mechanism. This enables the model to autonomously select query-relevant information across modalities, thereby addressing the challenges of cross-modal alignment and fine-grained understanding. Experimental evaluations across diverse benchmarks demonstrate that OmniAgent significantly outperforms existing open-source and closed-source OmniLLMs.

AI USE STATEMENT

We used generative AI tools to identify potential implementation errors in the evaluation code and assess the syntax, clarity, and readability of the manuscript. The authors manually reviewed and tested every code-related suggestion before accepting it. We did not use generative AI tools to develop theoretical models, formulate mathematical claims or proofs, design research methods, or determine the final interpretation of the results. The remaining disclosure requirements do not apply to this work. The authors assume full responsibility for the final content of this work, including any text, statements, and artifacts produced with the assistance of generative AI.

REFERENCES

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, et al. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*, 2025.

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*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Yuqin Cao, Xiongkuo Min, Yixuan Gao, Wei Sun, Zicheng Zhang, Jinliang Han, and Guangtao Zhai. Xgc-avis: Towards audio-visual content understanding with a multi-agent collaborative system. *arXiv preprint arXiv:2509.23251*, 2025.

Yijing Chen, Yihan Wu, Kaisi Guan, Yuchen Ren, Yuyue Wang, Ruihua Song, and Liyun Ru. Chronusomni: Improving time awareness of omni large language models. *arXiv preprint arXiv:2512.09841*, 2025.

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

Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Jun Chen, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Meerkat: Audio-visual large language model for grounding in space and time. In *ECCV*, 2024.

Sanjoy Chowdhury, Mohamed Elmoghany, Yohan Abeysinghe, Junjie Fei, Sayan Nag, Salman Khan, Mohamed Elhoseiny, and Dinesh Manocha. Magnet: A multi-agent framework for finding audio-visual needles by reasoning over multi-video haystacks. *arXiv preprint arXiv:2506.07016*, 2025.

Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. Mobilevlm: A fast, strong and open vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*, 2023a.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023b.

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.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025a.

Yang Ding, Yizhen Zhang, Xin Lai, Ruihang Chu, and Yujiu Yang. Videozoomer: Reinforcement-learned temporal focusing for long video reasoning. *arXiv preprint arXiv:2512.22315*, 2025b.

Wenjie Du, Li Jiang, Keda Tao, Xue Liu, and Huan Wang. Which heads matter for reasoning? rl-guided kv cache compression. *arXiv preprint arXiv:2510.08525*, 2025.

Baoyu Fan, Lu Liu, Xiaochuan Li, Runze Zhang, Liang Jin, and Jin Zhang. Fine-grained audio–visual event localization. *TNNLS*, 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 *ECCV*, 2024.

Sicheng Feng, Gongfan Fang, Xinyin Ma, and Xinchao Wang. Efficient reasoning models: A survey. *arXiv preprint arXiv:2504.10903*, 2025a.

Sicheng Feng, Kaiwen Tuo, Song Wang, Lingdong Kong, Jianke Zhu, and Huan Wang. Rewardmap: Tackling sparse rewards in fine-grained visual reasoning via multi-stage reinforcement learning. *arXiv preprint arXiv:2510.02240*, 2025b.

Sicheng Feng, Song Wang, Shuyi Ouyang, Lingdong Kong, Zikai Song, Jianke Zhu, Huan Wang, and Xinchao Wang. Can mllms guide me home? a benchmark study on fine-grained visual reasoning from transit maps. *arXiv preprint arXiv:2505.18675*, 2025c.

Siminfar Samakoush Galougah, Rishie Raj, Sanjoy Chowdhury, Sayan Nag, and Ramani Duraiswami. Aura: A fine-grained benchmark and decomposed metric for audio-visual reasoning. *arXiv preprint arXiv:2508.07470*, 2025.

Hong Gao, Yiming Bao, Xuezhen Tu, Yutong Xu, Yue Jin, Yiyang Mu, Bin Zhong, Linan Yue, and Min-Ling Zhang. Agentic video intelligence: A flexible framework for advanced video exploration and understanding. *arXiv preprint arXiv:2511.14446*, 2025.

Yuying Ge, Yixiao Ge, Chen Li, Teng Wang, Junfu Pu, Yizhuo Li, Lu Qiu, Jin Ma, Lisheng Duan, Xinyu Zuo, et al. Arc-hunyuan-video-7b: Structured video comprehension of real-world shorts. *arXiv preprint arXiv:2507.20939*, 2025.

Yuxin Guo, Shuailei Ma, Shijie Ma, Xiaoyi Bao, Chen-Wei Xie, Kecheng Zheng, Tingyu Weng, Siyang Sun, Yun Zheng, and Wei Zou. Aligned better, listen better for audio-visual large language models. In *ICLR*, 2025.

Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. In *ICLR*, 2016.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*, 2025.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *CVPR*, 2024.

Sullam Jeoung, Goeric Huybrechts, Bhavana Ganesh, Aram Galstyan, and Sravan Bodapati. Adaptive video understanding agent: Enhancing efficiency with dynamic frame sampling and feedback-driven reasoning. *arXiv preprint arXiv:2410.20252*, 2024.

Shixin Jiang, Jiafeng Liang, Jiyuan Wang, Xuan Dong, Heng Chang, Weijiang Yu, Jinhua Du, Ming Liu, and Bing Qin. From specific-mllms to omni-mllms: a survey on mllms aligned with multi-modalities. In *Findings of ACL*, 2025.

Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S Ryoo. Language repository for long video understanding. In *ACL*, 2025.

Noriyuki Kugo, Xiang Li, Zixin Li, Ashish Gupta, Arpandeep Khatua, Nidhish Jain, Chaitanya Patel, Yuta Kyuragi, Yasunori Ishii, Masamoto Tanabiki, et al. Videomultiagents: A multi-agent framework for video question answering. *arXiv preprint arXiv:2504.20091*, 2025.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Jiafu Tang, Zhenghao Song, Dingling Zhang, et al. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. *arXiv preprint arXiv:2510.10689*, 2025a.

Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, Song Chen, Xu Li, Da Pan, Shusen Zhang, Xin Wu, Zheng Liang, Jun Liu, Tao Zhang, Keer Lu, Yaqi Zhao, Yanjun Shen, Fan Yang, Kaicheng Yu, Tao Lin, Jianhua Xu, Zenan Zhou, and Weipeng Chen. Baichuan-omni technical report. *arXiv preprint arXiv:2410.08565*, 2024.

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

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *NeurIPS*, 2024.

Junan Lin, Daizong Liu, Xianke Chen, Xiaoye Qu, Xun Yang, Jixiang Zhu, Sanyuan Zhang, and Jianfeng Dong. Audio does matter: Importance-aware multi-granularity fusion for video moment retrieval. In *ACM MM*, 2025.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

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

Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment. *arXiv preprint arXiv:2502.04328*, 2025b.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *CVPR*, 2024.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. Drvideo: Document retrieval based long video understanding. In *CVPR*, 2025.

Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In *CVPR*, 2024.

OpenAI. Gpt-4 model, 2023. URL https://platform.openai.com. Accessed: 2023-11-08.

Ziqi Pang and Yu-Xiong Wang. Mr. video:" mapreduce" is the principle for long video understanding. *arXiv preprint arXiv:2504.16082*, 2025.

Jongwoo Park, Kanchana Ranasinghe, Kumara Kahatapitiya, Wonjeong Ryu, Donghyun Kim, and Michael S Ryoo. Too many frames, not all useful: Efficient strategies for long-form video qa. *arXiv preprint arXiv:2406.09396*, 2024.

Mengxue Qu, Xiaodong Chen, Wu Liu, Alicia Li, and Yao Zhao. Chatvtg: Video temporal grounding via chat with video dialogue large language models. In *CVPR*, 2024.

Kele Shao, Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Holitom: Holistic token merging for fast video large language models. *arXiv preprint arXiv:2505.21334*, 2025a.

Kele Shao, Keda Tao, Kejia Zhang, Sicheng Feng, Mu Cai, Yuzhang Shang, Haoxuan You, Can Qin, Yang Sui, and Huan Wang. When tokens talk too much: A survey of multimodal long-context token compression across images, videos, and audios. *arXiv preprint arXiv:2507.20198*, 2025b.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. In *ICML*, 2025.

Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. Audio-visual llm for video understanding. In *CVPR*, 2025.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Fine-grained audio-visual joint representations for multimodal large language models. *arXiv preprint arXiv:2310.05863*, 2023a.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-salmonn: Speech-enhanced audio-visual large language models. *arXiv preprint arXiv:2406.15704*, 2024.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*, 2023b.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn 2: Captioning-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025.

Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Dycoke: Dynamic compression of tokens for fast video large language models. In *CVPR*, 2025a.

Keda Tao, Kele Shao, Bohan Yu, Weiqiang Wang, Huan Wang, et al. Omnizip: Audio-guided dynamic token compression for fast omnimodal large language models. *arXiv preprint arXiv:2511.14582*, 2025b.

Keda Tao, Yuhua Zheng, Jia Xu, Wenjie Du, Kele Shao, Hesong Wang, Xueyi Chen, Xin Jin, Junhan Zhu, Bohan Yu, et al. Lvomnibench: Pioneering long audio-video understanding evaluation for omnimodal llms. *arXiv preprint arXiv:2603.19217*, 2026.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025a.

Meituan LongCat Team, Bairui Wang, Bin Xiao, Bo Zhang, Bolin Rong, Borun Chen, Chang Wan, Chao Zhang, Chen Huang, Chen Chen, et al. Longcat-flash-omni technical report. *arXiv preprint arXiv:2511.00279*, 2025b.

Qwen Team. Qwen3.8-omni: Towards native omni-modal agents. *arXiv preprint arXiv:2609.25611*, 2026a.

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

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. In *NeurIPS*, 2025.

Wenwen Tong, Hewei Guo, Dongchuan Ran, Jiangnan Chen, Jiefan Lu, Kaibin Wang, Keqiang Li, Xiaoxu Zhu, Jiakui Li, Kehan Li, et al. Interactiveomni: A unified omni-modal model for audio-visual multi-turn dialogue. *arXiv preprint arXiv:2510.13747*, 2025.

Mart Van Baalen, Andrey Kuzmin, Ivan Koryakovskiy, Markus Nagel, Peter Couperus, Cedric Bastoul, Eric Mahurin, Tijmen Blankevoort, and Paul Whatmough. Gptvq: The blessing of dimensionality for llm quantization. *arXiv preprint arXiv:2402.15319*, 2024.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

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

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

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *arXiv preprint arXiv:2503.13377*, 2025b.

Zikang Wang, Boyu Chen, Zhengrong Yue, Yi Wang, Yu Qiao, Limin Wang, and Yali Wang. Videochat-a1: Thinking with long videos by chain-of-shot reasoning. *arXiv preprint arXiv:2506.06097*, 2025c.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *CVPR*, 2025d.

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. *arXiv preprint arXiv:2512.05774*, 2025e.

Jianlong Wu, Wei Liu, Ye Liu, Meng Liu, Liqiang Nie, Zhouchen Lin, and Chang Wen Chen. A survey on video temporal grounding with multimodal large language model. *TPAMI*, 2025.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. Sheared llama: Accelerating language model pre-training via structured pruning. *arXiv preprint arXiv:2310.06694*, 2023.

Peiyu Xie, Jinxing Li, Guangming Lu, Yong Xu, and David Zhang. Caption assisted multimodal large language model for video moment retrieval. *TIP*, 2025.

Zhifei Xie and Changqiao Wu. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *arXiv preprint arXiv:2410.11190*, 2024.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

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

Haolin Yang, Feilong Tang, Lingxiao Zhao, Xiang An, Ming Hu, Huifa Li, Xinlin Zhuang, Yifan Lu, Xiaofeng Zhang, Abdalla Swikir, et al. Streamagent: Towards anticipatory agents for streaming video understanding. *arXiv preprint arXiv:2508.01875*, 2025a.

Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. Humanomniv2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 2025b.

Yudong Yang, Jimin Zhuang, Guangzhi Sun, Changli Tang, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, Zejun Ma, and Chao Zhang. Audio-centric video understanding benchmark without text shortcut. In *EMNLP*, 2025c.

Zeyuan Yang, Delin Chen, Xueyang Yu, Maohao Shen, and Chuang Gan. Vca: Video curious agent for long video understanding. In *CVPR*, 2025d.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *ICLR*, 2022.

Yufei Yin, Qianke Meng, Minghao Chen, Jiajun Ding, Zhenwei Shao, and Zhou Yu. Videoarm: Agentic reasoning over hierarchical memory for long-form video understanding. *arXiv preprint arXiv:2512.12360*, 2025.

Huaying Yuan, Zheng Liu, Junjie Zhou, Ji-Rong Wen, and Zhicheng Dou. Videodeepresearch: Long video understanding with agentic tool using. *arXiv preprint arXiv:2506.10821*, 2025.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, et al. Timesuite: Improving mllms for long video understanding via grounded tuning. *arXiv preprint arXiv:2410.19702*, 2024.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. In *EMNLP*, 2024a.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, Ying Shan, and Limin Wang. Timelens: Rethinking video temporal grounding with multimodal llms. *arXiv preprint arXiv:2512.14698*, 2025a.

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. *arXiv preprint arXiv:2406.16620*, 2024b.

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. *arXiv preprint arXiv:2505.18079*, 2025b.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024c. URL https://arxiv.org/abs/2410.02713.

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

Junhan Zhu, Hesong Wang, Mingluo Su, Zefang Wang, and Huan Wang. Obs-diff: Accurate pruning for diffusion models in one-shot. *arXiv preprint arXiv:2510.06751*, 2025a.

Muzhi Zhu, Hao Zhong, Canyu Zhao, Zongze Du, Zheng Huang, Mingyu Liu, Hao Chen, Cheng Zou, Jingdong Chen, Ming Yang, et al. Active-o3: Empowering multimodal large language models with active perception via grpo. *arXiv preprint arXiv:2505.21457*, 2025b.

Appendix Content

A. More Implementation Details P16
B. Behavior Analyses P16
B.1 Explainability of Modal Tendency Reasoning P16
B.2 Reasoning Behavior Analyses in Different Benchmarks P17
C. Efficiency Analyses P17
D. Case Study P18
D.1 Audio-Video Reasoning Cases P18
D.2 Unimodal-based Reasoning Cases P18
E. Future Work P19
F. ⭐ Insights and More Discussion P20
G. Prompts of OmniAgent P22
H. Prompts and Showcase of Toolset P23

A More Implementation Details

OmniAgent operates within an agentic, active reasoning paradigm. We employ the official APIs for all utilized tool models and the OpenAI model series. Conversely, for comparative baselines, we rely primarily on results reported by established benchmarks. When the reasoning process exceeds the maximum step limit, we leverage OpenAI o3 to synthesize answers based on accumulated information and evidence. To mitigate the impact of network fluctuations and latency, any queries interrupted by connectivity issues were re-evaluated in a final testing phase. For the DVD (Zhang et al., 2025b), we used the official code for evaluation and set the FPS to 5 for fair comparison. For our tools $\mathcal{T}_{\text{VGA}}$ and $\mathcal{T}_{\text{VCA}}$, we set the video sampling FPS to 2 and 5, respectively.

For DVD, due to the high API costs that we cannot afford for longer videos, we only conducted the test on Daily-Omni. For the Daily-Omni agent (Zhou et al., 2025), we adopt the official performance metrics reported in their paper. Due to the agent's exclusive reliance on the segmented video input configuration inherent to the Daily-Omni benchmark, cross-evaluation on other benchmarks was not feasible. Regarding tool integration, we incorporated a supplementary video metadata tool that enables the model to autonomously retrieve essential video attributes, including duration and frame rate (FPS), thereby providing foundational data to support subsequent tool invocations.

Cost. We measure the average API cost of our OmniAgent in three benchmarks. As the length of the video and the complexity of the query change, the total cost will also change accordingly. Thus, OmniAgent incurs a cost of \$0.05-\$0.11 per question with OpenAI, Qwen, and Gemini API.

Prompts. We show the prompts used by different tools and models within the proposed OmniAgent: (1) system prompts and user prompts for the agent (Section G); (2) prompts for the video tools (Sections H.1 and H.2); (3) prompts for the audio tools (Sections H.3, H.4 and H.6); (4) prompts for the event tools (Sections H.5 and H.7).

B Behavior Analyses

B.1 Explainability of Modal Tendency Reasoning

To analyze the performance of OmniAgent across diverse benchmarks: Daily-Omni (DO) (Zhou et al., 2025), WorldSense (WS) (Hong et al., 2025), and OmniVideoBench (OV) (Li et al., 2025a), we quantified the modality distribution of the tools utilized to derive answers. Our analysis focused exclusively on QA pairs correctly resolved by the agent. As illustrated in Figure 7, we categorized these instances into three primary classes based on input requirements: visual-only (video perception tools), audio-only (audio perception tools and event perception tools), and mix. We observe that OmniAgent can obtain a correct answer, depending on the usage of unimodal tools, on some questions

in the WorldSense benchmark, demonstrating the generalization ability of our agent (for more case studies, see Section D).

This indicates that during tool invocation, OmniAgent effectively infers the necessary context from the query and retrieves information aligned with the specific required modality. Furthermore, this demonstrates that our proposed algorithm is not strictly audio-dependent but rather possesses the capability to make autonomous, query-driven decisions. Driven by the specialized design of the Daily-Omni and OmniVideoBench, the agent is typically required to conduct a comprehensive analysis of both modalities before formulating a response. Therefore, OmniAgent has wide adaptability for omnimodal understanding. This also reflects that active perception can concentrate information in the required modalities.

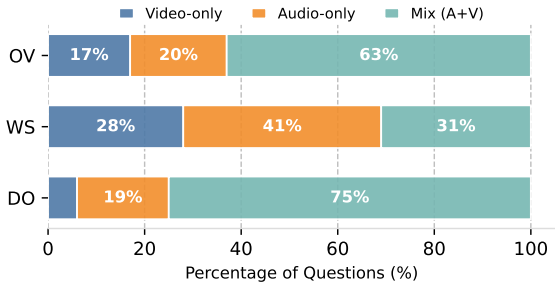

Figure 7: Focusing exclusively on the subset of queries correctly resolved by OmniAgent, we quantified the distribution of modal tools utilized during response generation.

B.2 Reasoning Behavior Analyses in Different Benchmarks

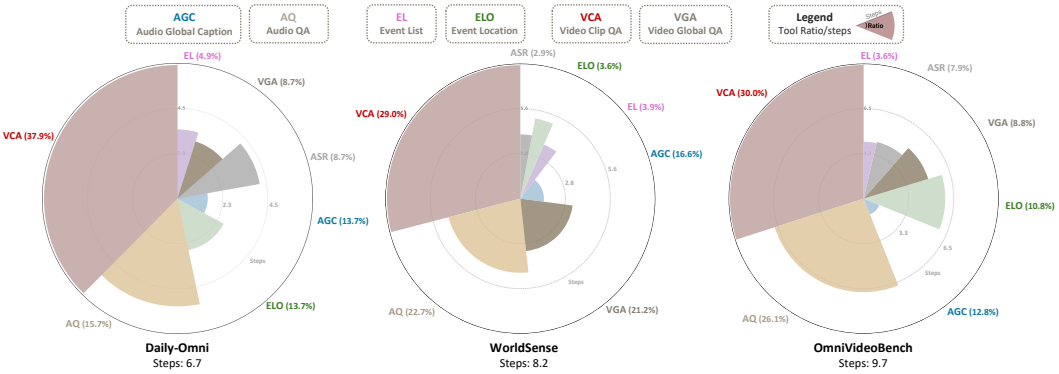

Figure 8: Analysis of the behavior of OmniAgent with OpenAI o3 in three different benchmarks.

Consistent with the methodology outlined in Section 4.3, OmniAgent adheres to a coarse-to-fine reasoning paradigm to achieve fine-grained understanding. Figure 8 presents the statistical analysis of inference behaviors across diverse benchmarks. Notably, while increased video duration correlates with a higher number of iterative inference steps for information acquisition, the overall distribution of tool usage remains stable. Regarding WorldSense, the dataset is characterized by a high prevalence of distinctly vision-based and audio-based questions. For vision-centric queries, the agent typically employs a "global-to-local" strategy (Figure 11), leading to increased invocation frequency for Video Global QA. Conversely, OmniVideoBench exhibits a balanced distribution between audio-based localization and coarse-grained video localization.

C Efficiency Analyses

As detailed in Table 5, a comparison with the DVD baseline (Zhang et al., 2025b) demonstrates that our method substantially reduces visual token redundancy. Moreover, despite the computational overhead incurred by processing audio signals, OmniAgent achieves a significant reduction in inference latency. Due to the prohibitive computational cost of DVD for the remaining benchmarks, we conducted our evaluation on a randomly sampled subset of queries. Notably, in scenarios with longer video durations,

Table 5: Comparison of token consumption and latency with DVD in the Daily-Omni benchmark. For a fair comparison, we use Qwen3-VL for the caption generation in DVD.

Method	**Input Token Cost** Visual Only	**Average Latency** On Daily-Omni Benchmark
DVD	18.6k	104s
OmniAgent (Ours)	8.3k	71s



Figure 9: **Complex reasoning case in long video.** For complex problems, the intelligent agent needs to conduct repeated iterations to find key evidence and information. Moreover, we have discovered that audio plays a crucial role throughout the reasoning process.

our approach reduces the aggregate API cost to approximately **10%** of the cost required by DVD. Notably, our design prioritizes solving the cross-modal alignment and fine-grained understanding over inference speed, and this work proposes a new paradigm for agent-based algorithms to solve audio-video understanding. We will work on more efficient improvements later.

D Case Study

D.1 Audio-Video Reasoning Cases

Figures 9 and 10 visualize the active decision-making and tool invocation mechanisms of OmniAgent when processing audio-visual queries. Figure 10 depicts the reasoning workflow for fundamental questions on short video clips. In this scenario, the agent leverages audio cues to temporally localize events before subsequently analyzing the corresponding visual information. In contrast, Figure 9 demonstrates a more complex reasoning scenario. The model initially establishes a global audio-visual context, performs joint audio-visual localization for the key cue "no signal", and integrates specific QA tools to extract decisive evidence. This exemplifies the adaptive operational logic of OmniAgent when addressing complex queries, highlighting the efficacy of audio-guided event localization.

D.2 Unimodal-based Reasoning Cases

As detailed in Section B.1, for tasks exhibiting a unimodal bias, such as those exclusively evaluating visual or audio comprehension, the agent invokes decision-making tools actively. Crucially, *OmniAgent does **not enforce strict** adherence to a specific video or audio modality; rather, modality selection is dynamically determined by the agent's internal reasoning.* As illustrated in Figure 11, when addressing visual queries, OmniAgent adopts a coarse-to-fine processing strategy, aligning with our design objective of facilitating fine-grained understanding. This demonstrates that while audio is

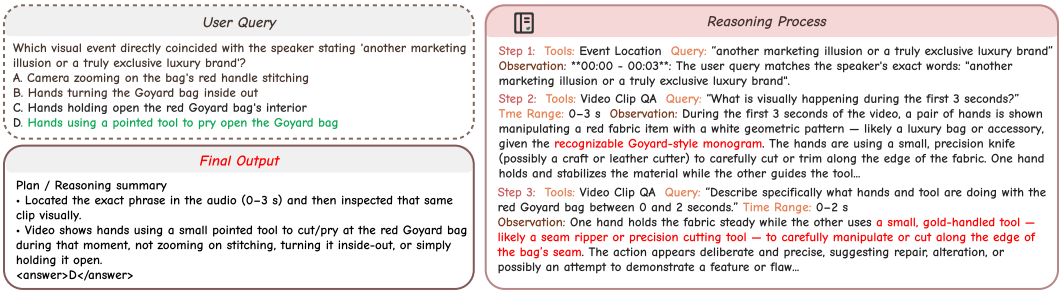

Figure 10: **Basic reasoning case in short video.** In tasks involving fundamental audio-visual event understanding, OmniAgent demonstrates high proficiency, deriving accurate answers within 2-3 reasoning steps. Specifically, the agent extracts temporal cues from the audio, subsequently leveraging this temporal grounding to align and analyze the corresponding visual information strictly.

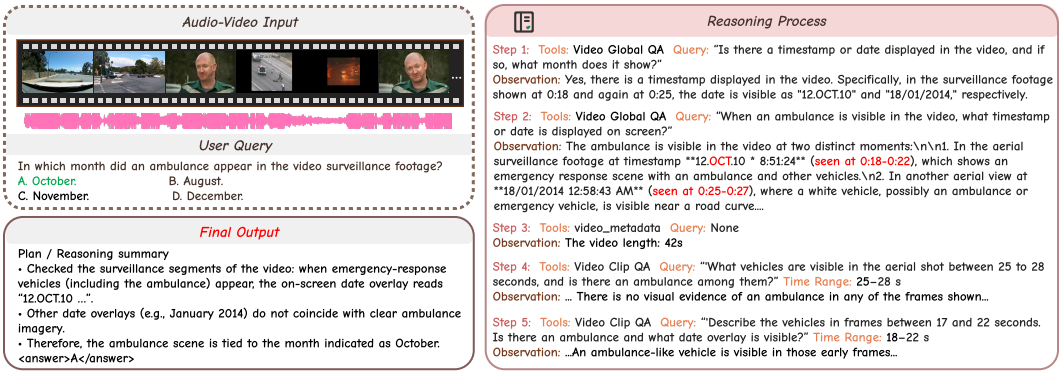

Figure 11: **Video-based reasoning process case.** For queries resolvable solely via visual information, OmniAgent autonomously elects to utilize video tools, adhering to a reasoning trajectory from coarse to fine. Typically, the agent initiates the process with Global QA for context retrieval and localization, subsequently employing Clip QA to verify fine-grained details. This validates that our framework is not rigidly constrained by audio dependency; instead, it employs adaptive, context-aware reasoning.

a powerful cue for localization, it is not forcibly integrated into scenarios where it is extraneous. Consequently, this confirms the capability of our agent to locate events solely via visual cues. Conversely, Figure 12 depicts a scenario where the solution is derived exclusively from audio information.

E Future Work

While our agent significantly advances omnimodal audio-video understanding, the reliance on iterative reasoning inevitably incurs higher computational overhead. Nevertheless, we maintain that orchestrating unimodal tools via agent reasoning is a promising direction for resolving current challenges in cross-modal alignment and fine-grained understanding. To enhance efficiency, future work will focus on training an Agentic omnimodal large language model with tool calling (Ding et al., 2025b; Tian et al., 2025). Concurrently, we observe that the active decision of the agent naturally introduces a degree of stochasticity. The error propagation remains a critical factor, as the accuracy of upstream tool outputs directly influences subsequent reasoning. Therefore, as we analyze in Section 4.4, Table 4, and Figure 12, the capability and output accuracy of the tool model are crucial for maintaining the correctness of the agent's reasoning. Moreover, given the substantial challenges inherent in deploying open-source models, particularly their significant inference overhead, optimizing the backbone model via acceleration and compression remains a critical pathway to enable the practical deployment of agents at scale (Han et al., 2016; Zhu et al., 2025a; Chu et al., 2023a; Lin et al., 2024; Du et al., 2025; Tao et al., 2025a; Shao et al., 2025a; Shen et al., 2025; Shao et al., 2025b; Van Baalen et al., 2024; Sun et al., 2023b; Xia et al., 2023; Feng et al., 2025a). Finally, how to polish

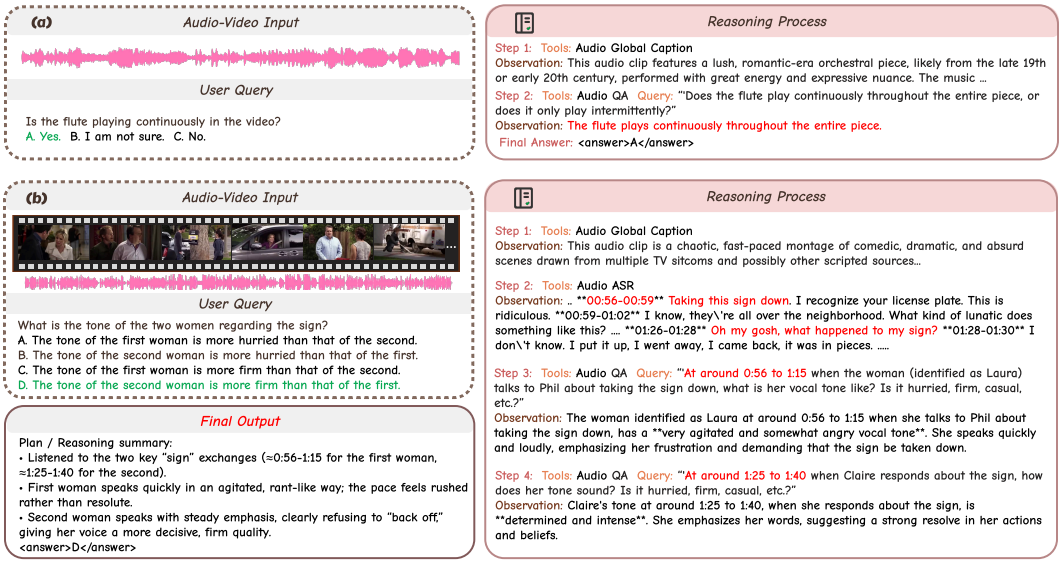

Figure 12: **Audio-based reasoning process case.** Regarding inquiries specifically targeting audio information within the video, OmniAgent strategically prioritizes audio tools, thereby avoiding superfluous visual processing and reducing information redundancy. For fundamental queries (a), the agent directly leverages audio-based interrogation to derive answers. In complex scenarios (b), the agent similarly adheres to a coarse-to-fine reasoning paradigm. As illustrated in the figure, the agent initially employs the ASR tool to transcribe the speaker's dialogue. It then analyzes this transcript relative to the query to isolate the relevant temporal segment, ultimately leveraging the Audio QA tool to verify fine-grained details, such as the speaker's emotional state. This workflow further validates the efficacy and logical soundness of our architectural design.

the tool outputs, build more efficient omnimodal memory, and integrate multi-agent frameworks constitutes key directions for our future research.

F ⭐ Insights and More Discussion

Currently, omnimodal understanding has emerged as a focal point of research. Driven by the intrinsic coupling of audio and visual modalities, this field is witnessing rapid advancement. However, constrained by data scarcity and architectural bottlenecks, existing end-to-end models often face the challenge of demonstrating robust, fine-grained cross-modal comprehension. Thus, in this work, we present a novel paradigm designed to address the complexities inherent in audio-video understanding. Drawing inspiration from human cognitive strategies for question answering, we introduce an active perception agent. Our approach not only achieves state-of-the-art results across multiple benchmarks but also significantly enhances the transparency of the reasoning process. Leveraging this interpretability, we facilitate a more profound discussion on multimodal understanding and offer critical insights derived from our experimental analysis.

Time Grounding Ability. OmniAgent effectively leverages the high information density and low redundancy intrinsic to audio signals to optimize event localization. However, as we analyze in Section 4.4, within the current research landscape, the temporal grounding capabilities of open-source MLLMs remain constrained. While prior research has predominantly prioritized temporal grounding within the visual domain (Wang et al., 2025b; Wu et al., 2025; Zhang et al., 2025a; Zeng et al., 2024; Qu et al., 2024; Huang et al., 2024), audio and audio-visual grounding have remained largely under-explored; consequently, advancing these capabilities constitutes a critical research imperative. Concurrently, numerous recent studies have leveraged post-training and RL to enhance performance in specific temporal grounding tasks; however, such targeted optimization may compromise the generalizability of models across broader domains. Consequently, achieving better temporal grounding within foundational MLLMs remains a formidable challenge.

Benchmark. As demonstrated in Section B.1, Figure 8, Figure 12, and Figure 11, there is a gap in existing evaluation datasets for audio-visual understanding. Curating a representative test dataset to evaluate the capacity of models for joint audio-visual comprehension presents significant challenges, yet it is of paramount importance. Consequently, we underscore the necessity for the future development of more rigorous and comprehensive benchmarks dedicated to holistic audio-visual understanding.

Audio-guided. In the realm of cross-modal reasoning, precise temporal alignment is paramount for accurate comprehension. This study underscores the pivotal contribution of audio to holistic multimodal understanding. As elaborated in Sections B and D.2, our agent transcends passive audio ingestion; instead, it actively arbitrates between auditory and visual information acquisition, adapting its strategy based on the query and the evolving reasoning process. Furthermore, we acknowledge the potential for temporal asynchrony between audio and video streams. Notably, OmniAgent demonstrates significant robustness in such scenarios, effectively mitigating the impact of misalignment. Concurrently, we strive to emulate human cognition by establishing a systematic framework that progressively validates evidence in a hypothesis-driven manner.

Tool Self-Calling of OmniLLMs. While the current reliance on external models and extended contexts improves performance, it constrains reasoning efficiency. To address this, in the future, we envision training an omnimodal agentic model. This architecture will ingest diverse modal inputs and feature tool self-calling, enabling the system to actively decide how to attend to specific audio or visual and address the bottleneck of inference cost.

Applicability for More Modal Tools. Our analysis of OmniAgent's reasoning patterns and empirical results reveals that orchestrating agents with single-modal tools offers a robust solution to the multimodal alignment challenge. Specifically, OmniAgent is capable of autonomously invoking tools via context-driven reasoning tailored to diverse problem settings. Consequently, the framework possesses inherent scalability, facilitating the future integration of additional modality-specific tools to address an expanded scope of multimodal understanding tasks.

G PROMPTS OF OMNIAGENT

G.1 AGENT SYSTEM PROMPT

You are the central reasoning brain of an audio–video analysis agent.
Your role:
- Answer the user's question about a given video by intelligently using the available tools (audio and video analysis).
- Follow a THINK → ACT → OBSERVE → REFLECT loop:
- THOUGHT: Reason step by step about what to do next.
- ACTION: Call exactly one tool that moves you closer to the answer.
- OBSERVATION: Read and interpret the tool's output, update your beliefs.
- REFLECTION: Reflect on the previous steps and the overall process.
General rules:
- Use both AUDIO and VIDEO information whenever they can help. Prefer to listen first, then look.
- Do not invent timestamps, file paths, or other arguments. Use values taken from the user input or from previous tool outputs.
- Be selective: tools may be noisy or incomplete. Cross-check and verify important information using multiple tools if needed.
- Stop calling tools once you have enough evidence to answer confidently.
Final answer style:
- When you are done with tools, reply directly to the user (no more tool calls).
- Start with a short "Plan / Reasoning summary" (1–3 bullet points) explaining briefly how you used audio vs video.
- Then give a clear, concise answer to the question.
- Do NOT expose raw tool-call traces or long chain-of-thought; keep the explanation high-level and user-friendly.

G.2 AGENT USER PROMPT

You are given a video and a question. Carefully read the question and think about how to combine AUDIO and VIDEO information to answer it.
Tool usage guidelines for this task:
- For a high-level understanding of the audio (topics, structure, key events), you can use audio_global_caption.
- For detailed questions about what is said or heard, you can use audio_qa and/or audio_ASR.
- When you care about WHEN things happen in the audio, prefer:
- Audio_EventList to get a rough timeline of major audio events.
- Audio_EventLocation to locate specific events or phrases by time.
- For visual understanding of the whole video, use video_global_qa.
- For fine-grained visual details in a short time range, use video_clip_qa.
If you need to choose or validate time ranges, call video_metadata to check the total duration and pick valid integer ranges.
Remember:
- Use audio to find time and content first whenever possible, then inspect the corresponding visuals: from listen to look.
- Plan your tool calls, but you are free to adjust the plan based on what you observe from previous tools.
Video path:
{video_path}
User question:
{question}

H PROMPTS AND SHOWCASE OF TOOLSET

H.1 VIDEO GLOBAL QA

You will be shown a video.
Your tasks are:
1. Carefully inspect what is happening in the video.
2. Reason step by step if necessary.
3. Answer the user's question as precisely as possible, always staying consistent with what is visible in the video.
User question:
{question}

H.2 VIDEO CLIP QA

You will be shown a video. Treat them as a short video: reason about how objects and people change over time across the frames, not just each image in isolation.
Your tasks:
1. Understand the main actions and changes that occur during this clip.
2. Reason step by step if needed.
3. Answer the user's question as precisely as possible, always staying consistent with what is visually supported in the frames.
You are analyzing a short VIDEO CLIP taken from a longer video.
1. It corresponds to the time range roughly from {start time} to {end time} seconds in the original video.
2. The frames are in temporal order and show how the scene evolves across this clip.
3. Assume the frames have ALREADY been correctly aligned with this time range; do not claim that they are from an earlier or later part of the video.
4. Answer ONLY about what happens within this clip. If the requested event is not visible here, say that it is not visible in this clip.
User question:
{question}

H.3 AUDIO ASR

You are a professional transcriber. Task: Generate a verbatim transcript of the speech, including precise timestamps for each sentence or natural segment.
REQUIRED OUTPUT FORMAT (Strict):
You must output a list where every line follows this exact format:
MM:SS-MM:SS Transcript text here
MM:SS-MM:SS Next sentence here
CRITICAL ANTI-REPETITION & NOISE RULES:
1. **Transcribe Speech Only**: Focus on clear spoken dialogue.
2. **Handle Repetitive Sounds**: If you hear repetitive noises (e.g., 'wu wu wu...', continuous laughter), **DO NOT** repeat the text. Instead, use a bracketed summary with the timestamp.

H.4 AUDIO GLOBAL CAPTION

Provide a high-level summary of the audio.
Focus on the main topics, key events, and the overall atmosphere.

H.5 EVENT LIST

You are an expert Audio Content Analyst. Your task is to generate a structured timeline of significant semantic events for the entire audio track.
Objective: Continuous & Semantic Timeline
Create a timeline that divides the audio into logical "chapters" or "scenes".
Coverage & Precision Rules (Critical):
1. Full Duration: You MUST start at 00:00 and cover the audio until the very end of the file. Do not skip any time periods.
2. No Gaps: The start time of a new segment should typically match (or be very close to) the end time of the previous segment.
The timeline must be contiguous. 3. Precise Boundaries: Listen carefully to identify the exact second where a scene transitions (e.g., when the music stops or a new speaker actually starts).
Segmentation Logic (Medium Granularity):
1. Merge, Don't Split: Treat a continuous conversation, a sustained musical piece, or a consistent environment as a SINGLE segment.
Example: If two people talk for 2 minutes about the same topic, that is ONE segment, not twenty short ones.
2. Trigger for New Segment: Only start a new segment when there is a definite shift in context:
- Topic change.
- Primary speaker switch (in formal structured turns, not quick banter).
- Distinct Speech-to-Music or Environment change.
3. Ignore Noise: Disregard short interruptions (<3s), coughs, or filler words.
Output Format:
1. Strictly output a Markdown bullet list.
2. Format: 'MM:SS - MM:SS: [Concise Description]'

This tool is designed to provide a time series explanation for the agent from an audio perspective when there are no clues. Here, we present an example of the output of this tool:

Showcase of Event List
* **00:00 - 00:19**: The audio begins with a suspenseful, percussive soundscape, featuring deep drums and a building, ominous atmosphere. This suggests a dramatic or tense opening.
* **00:19 - 00:25**: The previous soundscape fades out, transitioning into a calmer, more reflective musical score with a clear melody. This might indicate a shift in mood or a reflective moment.
* **00:25 - 00:29**: A male speaker begins to talk over the ongoing musical score, discussing the checkout process on Orbit and the numerous products involved. The music continues in the background.

H.6 AUDIO QA

You will be given one audio.
Your tasks are:
1. Carefully listen to the audio and understand what is being said and what sounds are present.
2. Reason step by step if necessary.
3. Answer the user's question as precisely as possible, using only what can be inferred from the audio.
User question:
{question}

H.7 Event Location

Role: Precision Audio Analyst.
Task: Locate the exact timestamps in the audio track that match the User Query.
User Query: {query}
Search Protocols (Strict):
1. Precision First: Pinpoint the exact second the event starts. Do not give vague ranges.
2. Point vs. Duration: For instant sounds (e.g., a gunshot, a scream), use a single timestamp: MM:SS.
For sustained events (e.g., a speech segment, a song), use a range: MM:SS - MM:SS.
3. Quantity Logic: - If the query specifies a count (e.g., "first time", "top 2", "last occurrence"), strictly obey it.
- If unspecified, list all clear occurrences (merge adjacent ones if $<$ 2s apart).
4. Anti-Hallucination: If the specific event is NOT found, output exactly: 'N/A: Event not found.'
Output Format:
1. Return a clean Markdown bullet list ONLY.
2. Format: 'Timestamp: [Context/Detail] Why this matches.'

I Impact Statement

OmniAgent seeks to advance the field of machine visual and auditory understanding, equipping artificial intelligence with superior precision and flexibility for comprehending complex audio-video input. By facilitating a paradigm shift from passive reception to active exploration, this approach offers robust cross-modal, fine-grained reasoning capabilities, which hold substantial social value. However, as this technology is designed for the in-depth analysis of real-world video and audio streams, it inherently necessitates the consideration of privacy protection and data sensitivity. We acknowledge these potential ethical implications. Consequently, our evaluation is strictly confined to open-source available benchmark datasets (Zhou et al., 2025; Hong et al., 2025; Li et al., 2025a). Furthermore, our system design prioritizes transparency and interpretability within the tool invocation and decision-making processes.