OmniSeek: Native Tool Integration for Multi-turn Audio-Visual Reasoning

Haibo Wang[1*] Jiteng Mu[2] Jialu Li[2] Jingru Yi[2]
Yuanjun Xiong[2] Jianming Zhang[2] Lifu Huang[1] Mingze Xu[2]
[1]University of California, Davis [2]Adobe Research
hibwang@ucdavis.edu; mingzex@adobe.com

Abstract

We present OmniSeek, an agentic framework that transforms an Omni Large Language Model (Omni-LLM) into an active, multi-turn reasoning agent with native tool use. Rather than passively processing an entire audio-visual sequence in a single forward pass, OmniSeek makes evidence acquisition part of the reasoning process: it dynamically decides whether to look or listen, and over which temporal window, to retrieve sparse but critical evidence across different modalities within long contexts. Through an iterative multi-turn protocol, the retrieved raw audio or visual segments are appended back into the context to support subsequent reasoning. To cold-start this capability, we build a data engine that synthesizes OmniTraj-170K, a corpus of multi-hop Chain-of-Thought trajectories with interleaved audio and visual evidence. We first supervise the model on these trajectories to instill multi-turn tool-use behavior, and then further optimize the policy via a two-stage reinforcement learning with verifiable rewards. Moreover, we introduce an Audio-Visual Necessity objective that explicitly rewards successful trajectories whose reasoning depends on both modalities, discouraging single-modality shortcuts. Extensive experiments across a wide range of benchmarks demonstrate that OmniSeek learns adaptive cross-modal evidence seeking and consistently improves audio-visual reasoning performance.

1. Introduction

Multimodal foundation models [1, 32, 38, 50, 51, 58, 65] are rapidly evolving toward Omni Large Language Models (Omni-LLMs) capable of jointly processing text, audio, and visual inputs [12, 14, 36, 47, 56, 57]. However, the prevailing paradigm in omni reasoning still relies on single-pass forward encoding, where the model passively ingests the entire audio-visual sequence at once. This design is convenient for short and information-dense inputs, but becomes limiting as the context grows longer, where fine-grained visual details and brief acoustic events can be easily diluted among large amounts of irrelevant content. Consequently, even when the necessary evidence is present in the input, the model may fail to isolate and compose it effectively, instead falling back on language priors or unimodal shortcuts.

Addressing this limitation requires more than extending textual reasoning over multi-modal inputs [7, 19, 34, 48]. The key missing capability is active evidence acquisition: rather than reasoning over a fixed perceptual context, the model should be able to let its evolving reasoning state guide what to inspect next. In particular, it should determine which modality is most informative, where in the sequence to search, and whether the currently accumulated evidence is sufficient. This is important for omni-modal reasoning where complementary audio and visual cues may need to be composed over multiple reasoning steps. We therefore formulate omni-modal reasoning as an agentic evidence-seeking process, in which perception and reasoning are interleaved through iterative, modality-specific retrieval.

To this end, we introduce ***OmniSeek***, a framework that transforms an Omni-LLM into an active, multi-turn audio-visual reasoning agent. As illustrated in Figure 1, OmniSeek operates through an iterative `<think>` → `<tool_call>` → `<observe>` loop [62], where the model can invoke specialized tools to retrieve targeted audio or video segments on demand. This *modality-decoupled* design allows the agent to flexibly route its attention across different modalities and temporal regions, for example, first locating an informative audio cue and then inspecting a different video segment for complementary visual evidence. Rather than only converting retrieved evidence into intermediate textual descriptions [7], OmniSeek directly appends the clipped raw audio or visual segments back into the model context. The model can therefore iteratively refine its evidence, determine when sufficient information has been collected, and synthesize multi-hop cross-modal evidence before producing the final answer.

We cold-start this complex behavior by constructing ***OmniTraj-170K***, a large-scale dataset comprising multi-turn Chain-of-Thought (CoT) trajectories with interleaved audio-

*Work done during an internship at Adobe Research.

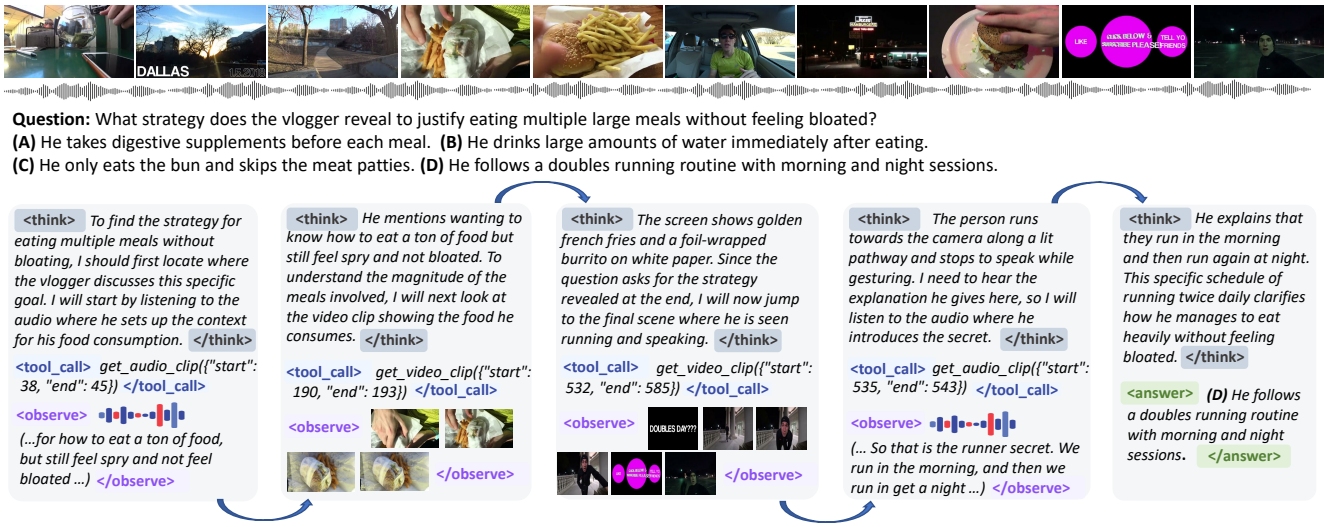

Figure 1. **Interleaved Multi-turn Audio-Visual Reasoning.** OmniSeek acts as an active agent operating through an iterative `<think>` → `<tool_call>` → `<observe>` loop. The agent dynamically alternates between fetching audio cues and extracting visual evidence from different time spans to answer a complex multi-hop question, avoiding the pitfalls of single-modality shortcuts.

visual evidence. To comprehensively cover real-world scenarios, this corpus spans diverse question types that demand explicit audio-visual integration. Specifically, each trajectory is grounded in fine-grained, timestamped video and audio segments, guiding the model to seamlessly interleave and reason over mixed modalities. Building upon this, we further optimize the policy via two-stage reinforcement learning (RL). Moreover, to avoid inadvertently reinforcing single-modality shortcuts during RL training, we introduce an **Audio-Visual Necessity** objective, which leverages modality-specific attention masking to credit reasoning trajectories that depend on evidence from both modalities, discouraging hallucinated grounding without requiring additional rollouts.

In summary, our contributions are as follows: *(1)* We propose OmniSeek, a novel agentic framework that enables Omni-LLMs to retrieve, interleave, and reason over decoupled audio-visual evidence across multiple turns. *(2)* We construct OmniTraj-170K, a dataset of multi-turn CoT trajectories with interleaved audio-visual evidence. *(3)* We design an Audio-Visual Necessity objective during RL training, which rewards trajectories that exhibit dependence on both modalities and discourages single-modality shortcuts. *(4)* Extensive experiments demonstrate that OmniSeek learns adaptive reasoning behavior and achieves leading performance across a broad suite of audio-visual benchmarks.

2. Related Work

Thinking with Images/Videos. The success of Chain-of-Thought (CoT) [53] in Large Language Models has inspired recent efforts to extend explicit reasoning processes into the multimodal domain [17, 19, 23, 29, 31, 34, 59]. Early explorations focused on image-level perception. For instance, DeepEyes [69] and Pixel Reasoner [41] incentivized models to "think with images" by natively invoking pixel-space operations (e.g., zoom-in, crop) via reinforcement learning, thereby shifting from passive global perception to proactive visual inspection. DeepEyesV2 [26] further broadened this agentic paradigm by incorporating external tools like code execution and web search. While these works successfully established active spatial exploration, extending this paradigm to the temporal dimension introduces distinct challenges. Frameworks such as Video-o3 [64], VITAL [66], LongVT [61], and VideoZoomer [16] transform video understanding into a multi-turn, global-to-local retrieval process with grounding capabilities [27, 39, 49]. These methods equip models with temporal zooming tools, enabling them to fetch and inspect high-frame-rate clips on demand. While most of these methods rely on explicit tool interaction, Open-o3-Video [37] instead embeds spatio-temporal coordinates directly into the reasoning trace for improved grounding.

Omni Large Language Models (Omni-LLMs). The evolution of multimodal foundation models has rapidly advanced toward the capability of jointly processing text, images, video, and audio [20, 28, 36, 67]. Previous works such as VideoLLaMA-2 [9] and Video-SALMONN 2 [43] have demonstrated significant improvements in audio-visual question answering. Recent Omni-LLMs, including Nemotron-3-Omni [14], the Qwen-Omni series [47, 56, 57], and MiniCPM-o-4.5 [12], further unify perception and generation across modalities while preserving strong unimodal capabilities. They also support longer contexts and finer-grained audio-visual grounding. Alongside advances in these foundation models, the high computational overhead of processing long audio-visual sequences has spurred research into efficient omni-modal inference, with works such as OmniZip [44], OmniSIFT [15], and OmniPack [42], which have

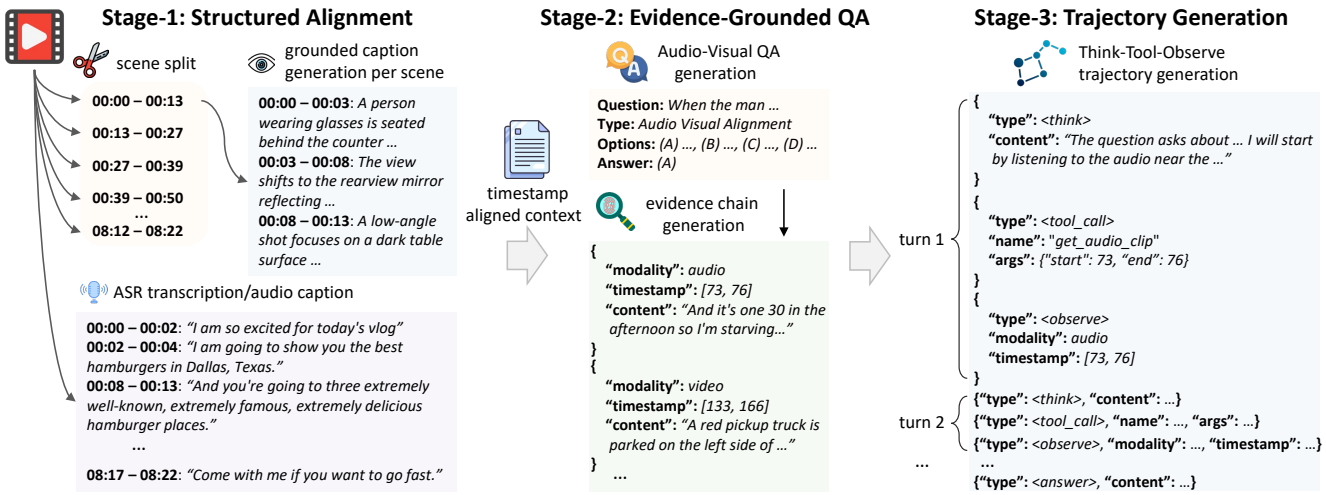

Figure 2. **Overview of OmniTraj-170K data engine.** The pipeline extracts timestamp-aligned audio-visual contexts (Stage-1), synthesizes questions grounded in strict cross-modal evidence chains (Stage-2), and translates them into multi-turn reasoning trajectories (Stage-3).

introduced diverse token compression strategies to accelerate inference and reduce memory footprints. Beyond efficiency, while methods like LatentOmni [13], OmniVideo-R1 [7], and OmniReasoner [6] explore omni-modal reasoning, they primarily rely on text-based reasoning traces, coupled modality, or single-turn retrieval. ***OmniSeek*** instead overcomes these bottlenecks by employing an iterative, multi-turn tool protocol for dynamic, independent modality routing, powered by a data engine that synthesizes reasoning trajectories with interleaved audio-visual evidence.

3. OmniTraj-170K

We design a three-stage data engine to generate high-quality demonstrations for policy warm-starting (Figure 2). The resulting ***OmniTraj-170K*** contains ∼170K audio-visual trajectories involving multi-turn, multi-hop reasoning.

Stage-1: Structured Audio-Visual Alignment. To establish timestamp alignment between modalities, we adopt a decoupled, parallel audio-visual annotation strategy. For the visual track, we first utilize PySceneDetect [3] to identify visual transitions and segment raw videos into multiple fine-grained shots. To avoid truncating ongoing actions, we sequentially merge these adjacent shots until their accumulated duration reaches a contextual window of approximately ∼15 seconds to form a cohesive scene, thereby preserving natural semantic boundaries. Operating scene-by-scene, Qwen3.5-397B-A17B [38] then generates a series of dense, timestamped visual captions for each scene. Processing within these short, ∼15-second-level contextual windows facilitates detailed and high-quality grounded captions. We strictly constrain the model to describe only visible actions, objects, and on-screen text, explicitly forbidding external knowledge or inferring content not visible in the scene. Finally, to resolve coreference and maintain entity consistency across scenes, the model additionally generates a detailed global caption for the entire video. Concurrently for the audio track, we acquire timestamped speech information directly from the source video's original automatic speech recognition (ASR) transcripts. Alternatively, specialized audio models such as Qwen3-Omni-Captioner [57] or Qwen3-ASR [40] can be deployed to extract dense, timestamped audio captions for each scene. This parallel pipeline yields a comprehensive, dual-track aligned context where all visual and auditory events are deterministically anchored to absolute timestamps.

Stage-2: Evidence-Grounded QA Generation. To prevent task homogenization and single-modality shortcuts, we utilize the aligned dual-track context to synthesize 19 types of audio-visual questions, where we explicitly instruct the Qwen3.5-397B-A17B [38] to consider both modalities and generate questions that cannot be answered without either audio or video cues. Crucially, rather than producing isolated question-answer pairs, we additionally instruct Qwen3.5-397B-A17B to plan a structured *evidence chain* required to arrive at the correct answer (see Figure 2, Stage-2). Each question is strictly associated with an ordered list of 2 to 7 evidence spans. To reduce annotation hallucination, the textual content and absolute timestamps for each span are copied directly from the Stage 1 annotations. Each span specifies its required modality (audio or video), the exact timestamp window, and its textual content. Notably, these spans are arranged in a logical retrieval sequence rather than strictly chronological order, mimicking the analytical multi-hop process of a human solver. We enforce structural constraints during the QA generation: every evidence chain must explicitly contain at least one audio span and one video span. This constraint is designed to reduce questions that can be answered from language priors or a single modality alone.

Stage-3: Interleaved Trajectory Assembly. At last, we

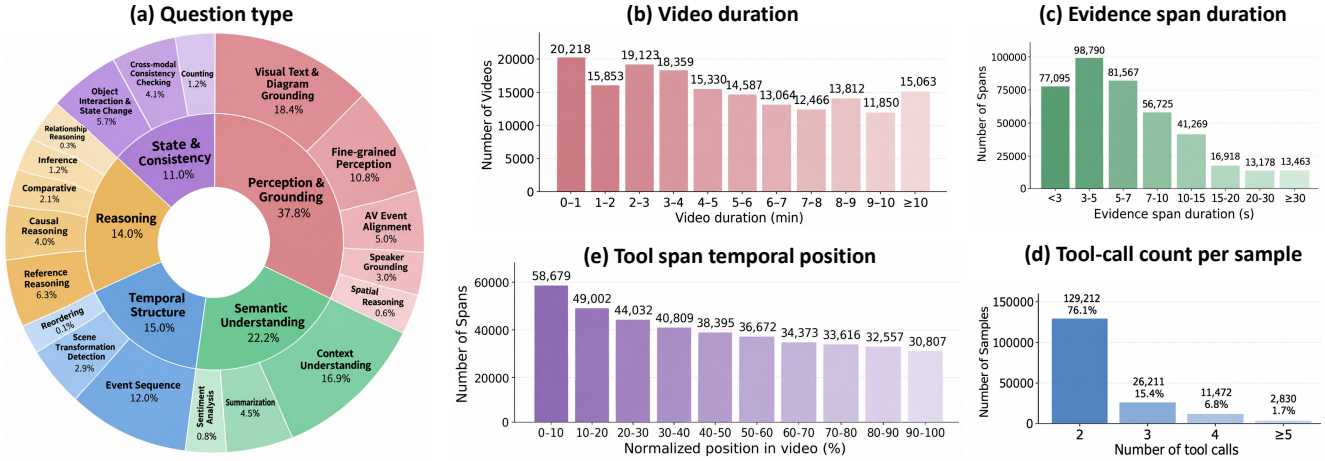

Figure 3. **OmniTraj-170K Statistics:** The dataset contains 169,725 multi-turn trajectories over 39,797 videos and covers 19 cross-modal question types. We report the distributions of (a) question types, (b) source-video durations, (c) evidence-span durations, (d) tool calls per trajectory, and (e) normalized temporal positions of retrieved evidence spans.

translate the evidence-grounded questions into multi-turn reasoning trajectories [62] as in Figure 2, Stage-3. A target trajectory operates through an iterative `<think>` → `<tool_call>` → `<observe>` loop, culminating in a final `<answer>`. To construct this complex sequence without suffering from hallucinations, we decouple the generation of the reasoning traces `<think>` from the tool execution step `<tool_call>` and returned observations `<observe>`. Specifically, the tool invocations (e.g., `get_video_clip` or `get_audio_clip` with exact timestamp arguments) and their corresponding observation contents are deterministically constructed from the `modality` and `timestamps` of the previously generated evidence chain. The Qwen3.5-397B-A17B is tasked only with generating the internal `<think>` nodes to bridge these predefined actions and observations. We instruct the model that the initial `<think>` node plans the retrieval, subsequent nodes reflect on the retrieved clips to guide the next hop, and the final node synthesizes the previous evidence before outputting the answer. Finally, our pipeline interleaves the model-generated `<think>` steps with the deterministically constructed `<tool_call>` and `<observe>` steps. We then verify that every trajectory contains only valid tool calls consistent with the predefined evidence spans and preserves the intended cross-modal evidence structure.

Dataset Statistics. Figure 3 summarizes the key statistics of OmniTraj-170K, with 169,725 trajectories over 39,797 videos. The corpus features a diverse distribution across 19 cross-modal question types. The source videos, derived from the FineVideo [18], span 122 diverse categories (e.g., education, science, news, and sports) and vary in length, ranging from under one minute to tens of minutes. Meanwhile, the extracted evidence spans are tightly localized, with the vast majority lasting between 3 and 10 seconds. In terms of multi-hop complexity, 76.1% of the trajectories require two tool calls, while the remaining 23.9% require three or more tool invocations to synthesize the final answer. The temporal positions of these retrieved spans are distributed across the entire video, exposing the model to retrieval targets at both early and late temporal positions.

4. OmniSeek

4.1. Multi-turn Tool Protocol

Given a video V alongside its synchronized audio stream A, the model initially ingests the full sequence as a coarsely sampled global context. However, rather than passively relying on this diluted context, OmniSeek operates under an iterative agent loop [62]. At each turn t, the model navigates through a structured state machine: `<think>` → `<tool_call>` → `<observe>`. Crucially, the actual modality content retrieved by the executed tools is dynamically appended back into the ongoing context as an `<observe>` node, continuously enriching the model's working memory with new sensory evidence. This active paradigm endows the model with the autonomy to dynamically route its attention, allowing the model to augment the initial global context through iterative evidence retrieval. To support flexible and precise evidence retrieval, OmniSeek decouples tool invocation across modalities, equipping the agent with two core operations:

- `get_audio_clip(start, end)`: Directs the agent to temporally isolate and inspect a targeted audio segment within a specific time window (`start`, `end`).
- `get_video_clip(start, end, fps, resolution)`: The agent dynamically determines the temporal boundaries of the targeted segment. Upon invocation, the tool automatically retrieves this local clip at a higher, self-defined sampling frame rate `fps` and spatial `resolution` compared to the coarse global

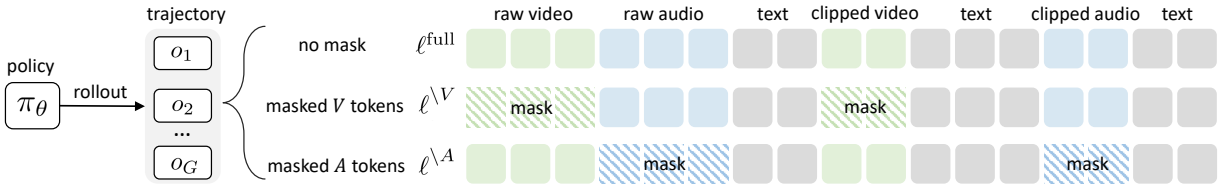

Figure 4. **Audio-Visual Necessity.** We measure audio-visual dependence through modality-specific attention masking.

input, facilitating a coarse-to-fine visual examination.

This asynchronous tool design empowers OmniSeek to break away from rigid modality-coupled constraints [6]. For instance, the agent can first capture an audio cue, and subsequently search an entirely different video segment with higher resolution to capture fine-grained visual details, achieving a coarse-to-fine inspection. The termination condition of this multi-turn loop is not hardcoded. Instead, it relies on the model's dynamic self-reflection. During the <think> process at each turn t, the model evaluates its current information state and assesses whether the accumulated cross-modal evidence chain within its context is sufficient to derive the answer. If further inspection is needed, it plans the next <tool_call>; if the evidence is conclusive, it terminates the loop and outputs the final <answer>.

4.2. Three-Phase Training Strategy

While the OmniTraj-170K dataset provides high-quality demonstrations of cross-modal reasoning, relying solely on behavioral cloning can lead to policy degradation [10, 30], where the model mimics the tool-calling format but implicitly falls back on single-modality shortcuts. Therefore, we design a progressive, three-phase training pipeline encompassing cold-start supervised fine-tuning (SFT) and verifiable-reward reinforcement learning (RL).

Phase 1: Cold-Start via Supervised Fine-Tuning. We first supervised fine-tune the base Omni-LLM on our OmniTraj-170K. The objective at this stage is primarily format and behavioral alignment: instilling the <think> → <tool_call> → <observe> syntax and warming up the model's ability to route its attention across interleaved audio and video tokens over multiple turns.

Phase 2: Broad Exploration via RL. Once the model has learned the multi-turn protocol, RL can be conducted on datasets with verifiable answers without annotated reasoning trajectories. Specifically, we continue training using Group Sequence Policy Optimization (GSPO) [68] on a subset of 30K multiple-choice questions from OmniVideo100K [2] and an additional 1K samples from the VideoHolmes [8] training split. Because they lack ground-truth reasoning trajectories, the agent must autonomously explore the environment to discover effective tool-use strategies. To guide this, we define three rule-based rewards. First, the **Accuracy Reward (r_{acc})** is a binary score assessing if the final predicted <answer> exactly matches the ground truth. Second, the **Format Reward (r_{format})** penalizes trajectories that violate the required structural tags, such as missing <think> closures. Finally, we introduce a **Tool-Use Reward (r_{tool})** to encourage the model to use retrieval tools rather than relying solely on the initial context. This reward is granted *only* if the agent successfully invokes at least one tool and ultimately answers the question correctly, which encourages successful tool use while avoiding credit for tool calls in incorrect trajectories. The overall reward for each generated trajectory during this phase is computed as the sum of these three components: $R = r_{\text{acc}} + r_{\text{format}} + r_{\text{tool}}$.

Phase 3: Hard-Example Refinement. In the final phase, we re-evaluate the Phase 2 model checkpoints on the OmniTraj-170K dataset to mine failure cases. From these instances, we construct a class-balanced subset of 8K hard examples where the model previously failed, which typically feature strong modality interference or demand complex multi-hop reasoning. We resume GSPO training on this challenging subset but employ a larger rollout size G to enable broader exploration during RL. To suppress single-modality shortcuts on these difficult questions, we additionally introduce an **Audio-Visual Necessity Reward (r_{avn})**, which we detail next in Sec. 4.3. This reward is designed to penalize trajectories that arrive at the correct answer while exhibiting weak dependence on one of the modalities, encouraging the generated trajectory to depend on both visual and auditory evidence. The overall reward in this final phase is thus: $R = r_{\text{acc}} + r_{\text{format}} + r_{\text{tool}} + r_{\text{avn}}$.

4.3. Audio-Visual Necessity

The accuracy reward r_{acc} is outcome-oriented and blind to the underlying reasoning process. A trajectory relying on both modalities receives the same credit as one exploiting single-modality shortcuts. To prevent the model from learning "hallucinated grounding" without actually attending to both streams, we introduce the **Audio-Visual Necessity Reward** (r_{avn}). This objective provides a token-level proxy for the trajectory's dependence on audio and visual context, shaping the policy without overriding the accuracy objective.

Necessity via Attention Masking. We measure necessity through specific attention masking on the model's generated rollout, avoiding the cost of re-generation. Given a sampled trajectory τ composed of tokens s_t, and the set of model-generated tokens $\mathcal{M}$ (i.e., response tokens within <think> and <answer> tags), let A and V denote the sets of all audio and visual tokens in the context. Under standard generation, the log-likelihood of emitting token s_t is

Model	Size	Daily-Omni (44s)	AVUT (69s)	WorldSense (141s)	FutureOmni (166s)	OmniVideoTest (168s)	VideoHolmes (184s)	JointAV (212s)	OmniVideoBench (409s)	MMOU (757s)	LVOmni (2049s)
Gemini-3.0-Pro [46]	-	81.1	-	66.4	-	-	67.0	-	61.8	-	65.8
Gemini-2.0-Flash [11]	-	67.8	-	56.2	-	-	30.6	-	41.5	-	42.9
Qwen3.5-Omni-Flash [47]	-	81.8	81.4	57.9	-	-	57.3	-	-	-	-
VideoLLaMA2 [9]	7B	35.2	44.9	25.4	40.8	-	35.2	46.8	29.2	28.4	27.2
VITA-1.5 [22]	7B	52.6	-	36.9	48.7	41.0	-	-	36.4	-	-
Qwen2.5-Omni [56]	7B	62.1	-	45.4	38.9	42.8	16.4	56.5	36.5	31.3†	32.0
Uni-MoE-2.0-Omni [35]	30B	64.3	-	-	52.8	46.9	-	-	38.6	-	-
OmniReasoner [6]	7B	64.2	-	46.7	-	-	40.0	-	34.8	-	35.4
OmniAgent [55]	7B	64.8	-	47.2	-	-	-	-	37.1	-	-
OmniVinci [63]	7B	66.5	-	48.2	-	-	-	-	-	27.8	-
LatentOmni [13]	7B	67.4	-	48.9	-	-	-	-	35.4	-	35.1
Qwen3-Omni-Instruct [57]	30B	71.9	76.5	55.1	53.6	54.5	59.1	63.6	43.6	54.1	35.8
Qwen3-Omni-Thinking [57]	30B	73.6	71.7	52.7	50.8	50.7	57.3	63.4	39.6	53.8	31.9
video-SALMONN 2+ [43]	72B/7B	79.4	72.2	56.5	47.0	45.2	57.8	46.7	36.7	-	32.7
Nemotron-3-Omni [14]	30B	74.5	-	55.2	-	-	-	-	-	-	-
OmniVideo [2]	30B	76.6	-	-	57.6	63.6	-	-	44.8	-	-
MiniCPM-o 4.5 [12]	9B	80.2	78.6	55.7	56.1	-	64.3	60.0	-	46.8†	34.8
OmniVideo-R1 [7]	30B	**82.8**	-	**65.8**	-	-	62.9	-	44.8	-	-
OmniSeek (ours)	30B	80.0	**78.8**	62.4	**58.3**	**69.5**	**74.6**	**72.8**	**47.7**	**70.4**	**44.2**

Table 1. Performance comparison of different methods on audio-visual benchmarks. Best open-source results are highlighted in bold.

$\ell_t^{\text{full}} = \log \pi_\theta(s_t \mid s_{<t}, V, A)$. Keeping τ fixed, we perform two auxiliary teacher-forced forward passes. As in Figure 4, in each pass, we only ablate one modality by zeroing out its keys in the attention mask ($\varnothing_A$ or $\varnothing_V$), leaving the remaining sequence unchanged. This intervention reduces the distribution shifts caused by feature replacement or removal, yielding the counterfactual log-likelihoods $\ell_t^{\setminus A} = \log \pi_\theta(s_t \mid s_{<t}, V, \varnothing_A)$ and $\ell_t^{\setminus V} = \log \pi_\theta(s_t \mid s_{<t}, \varnothing_V, A)$. For each response token $t \in \mathcal{M}$, we compute the drop in log-likelihood compared to the full-context pass to define the per-token necessity of each modality:

$$\Delta_t^A = \ell_t^{\text{full}} - \ell_t^{\setminus A}, \quad \Delta_t^V = \ell_t^{\text{full}} - \ell_t^{\setminus V}. \tag{1}$$

We then aggregate these token-level drops using a rectification function $[x]_+ = \max(x, 0)$, since negative values typically reflect token competition rather than anti-grounding, and near-zero values correspond to modality-agnostic template tokens. The modality-specific necessities are defined:

$$\text{nec}_A = \frac{1}{|\mathcal{M}|} \sum_{t \in \mathcal{M}} \left[\Delta_t^A\right]_+, \quad \text{nec}_V = \frac{1}{|\mathcal{M}|} \sum_{t \in \mathcal{M}} \left[\Delta_t^V\right]_+. \tag{2}$$

Finally, we formulate the overall necessity reward as a logical conjunction. Because our objective is to reward joint dependence on both modalities, simply averaging the drops would incorrectly allow a strong single-modality reliance to compensate for a near-zero reliance on the other. Instead, we apply the $\min$ operator to act as a logical AND gate, crediting the trajectory strictly by its *weaker* modality:

$$r_{\text{avn}} = c \cdot \min\big(\text{nec}_A,\ \text{nec}_V\big), \tag{3}$$

where $c \in \{0, 1\}$ indicates the correctness of the answer. This gating ensures we only reinforce joint modality dependence on successful trajectories. This is also computationally efficient: since ℓ_t^{full} is already obtained during the initial policy-gradient pass, computing r_{avn} only requires two lightweight `no_grad` passes without additional rollouts.

5. Experiments

Implementation Details. We initialize OmniSeek using Qwen3-Omni-30B-A3B-Instruct [57] and train it following the three-phase strategy outlined in Section 4.2. In Phase 1 (Cold-Start SFT), to preserve the base model's generalization, we apply a mixed-data strategy. Specifically, only 10% of the training samples utilize the interleaved tool-calling trajectories from OmniTraj-170K, while the remaining 90% consist of standard, single-turn QA settings. In Phase 2, we optimize the policy using the GSPO [68] algorithm with a learning rate of 1×10^{-6}. We set the rollout size to $G = 8$ sampled trajectories per prompt. In Phase 3, we resume training from the Phase 2 checkpoint, but we double the rollout size to $G = 16$, guided by the full reward formulation including r_{avn}. More details are provided in Appendix **??**.

Evaluation Benchmarks. To validate the effectiveness of OmniSeek, we conduct comprehensive evaluations across 10 omni benchmarks: Daily-Omni [71], AVUT (Human) [60], WorldSense [25], FutureOmni [5], OmniVideoTest [2], VideoHolmes [8], JointAVBench [4], OmniVideoBench [33], MMOU (test-mini) [24], and LVOmniBench [45]. Furthermore, to ensure OmniSeek preserves the model's foundational perception capabilities on broader scenes, we also evaluate it on 4 general video understanding benchmarks: Video-MME [21], LongVideoBench [54], MLVU [70], and LVBench [52]. See more details in Appendix A.2.

5.1. Main results

Omnimodal Understanding. As in Table 1, OmniSeek delivers state-of-the-art or highly competitive performance among open-source models across a broad suite of omnimodal benchmarks. These gains are most pronounced in

Model / Variant	P1	P2	P3	r_{avn}	Daily-Omni	WorldSense	FutureOmni	OmniVideoTest	VideoHolmes	OmniVideoBench	LVOmni	Video-MME
(a) Base Model	–	–	–	–	71.9	55.1	53.6	54.5	59.1	43.6	35.8	76.8
(b) + Phase 1 SFT	✓	–	–	–	69.1	50.3	50.1	55.8	55.9	38.4	35.0	77.3
(c) + Phase 2 RL	✓	✓	–	–	75.5	55.0	57.2	63.0	69.6	45.2	41.6	76.8
(d) + Phase 3 RL (G=8)	✓	✓	✓	–	75.9	58.6	56.4	65.9	72.6	46.8	43.4	77.7
(e) + Phase 3 RL (G=16)	✓	✓	✓	–	78.2	59.2	58.1	67.3	71.7	47.0	43.6	77.9
(f) + Phase 3 RL (G=16)	✓	✓	✓	✓	80.0	62.4	58.3	69.5	74.6	47.7	44.2	78.5
Δ vs. (a) Base	–	–	–	–	+8.1	+7.3	+4.7	+15.0	+15.5	+4.1	+8.4	+1.7
Δ vs. (e) w/o r_{avn}	–	–	–	–	+1.8	+3.2	+0.2	+2.2	+2.9	+0.7	+0.6	+0.6

Table 2. Ablation study on training strategy. P1, P2, and P3 denote Phase-1 SFT, Phase-2 RL, and Phase-3 RL; G is the rollout size.

Model	Size	Video-MME (w/o sub, 1059s)	LongVideoBench (730s)	MLVU (m-avg, 705s)	LVBench (4038s)
Gemini-3.0-Pro [46]	-	88.6	75.9	75.7	77.0
Gemini-2.0-Flash [11]	-	72.4	-	71.0	57.9
Qwen3.5-Omni-Flash [47]	-	77.0	-	81.9	65.7
Visual-only inputs					
SlowFast-LLaVA-1.5 [58]	7B	63.9	62.5	71.5	45.3
LongVT [61]	7B	64.3	-	-	41.3
Video-Zoomer [16]	7B	64.6	55.9	69.9	44.0
VideoLLaMA3 [65]	7B	66.2	59.8	73.0	45.3
LLaVA-OneVision [32]	72B	66.2	61.3	66.4	-
Video-o3 [64]	7B	66.5	60.5	72.1	47.6
InternVL-3.5 [51]	30B	68.7	63.8	73.0	-
Audio-Visual inputs					
OmniAgent [55]	7B	67.8	-	71.1	50.5
MiniCPM-o 4.5 [12]	9B	70.4	66.0	76.5	50.9
video-SALMONN 2+ [43]	7B	73.4	-	73.6	49.7
OmniVideo-R1 [7]	30B	73.6	-	74.1	**51.9**
Qwen3-Omni-Thinking [57]	30B	74.3	-	72.9	49.0
Qwen3-Omni-Instruct [57]	30B	76.8	-	75.2	50.2
OmniSeek (ours)	30B	**78.5**	**66.4**	**77.1**	51.4

Table 3. Performance on general video benchmarks.

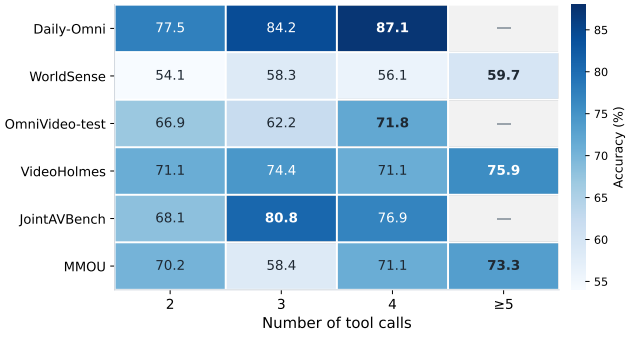

Figure 5. Accuracy vs. Number of tool calls.

long-form and complex scenarios. On benchmarks featuring long-form video lengths, such as MMOU and LVOmni, OmniSeek achieves **70.4%** and **44.2%**, outperforming the base Qwen3-Omni-Instruct by margins of +16.3% and +8.4%, respectively. Similarly, on VideoHolmes, OmniVideoTest and OmniVideoBench, where visual clues are sparsely scattered and require proactive inspection, OmniSeek leads the second-best open-source competitors by significant margins. These results are consistent with the hypothesis that OmniSeek's active `get_video_clip` and `get_audio_clip` tool-use paradigm can successfully isolate high-resolution visual and audio evidence on demand, and mitigate information loss in long audio-visual contexts. Furthermore, these results highlight the advantage of active tool-use over pure textual Chain-of-Thought (CoT) models like OmniVideo-R1. While such text-only reasoning models perform strongly on benchmarks with shorter, denser contexts, they encounter bottlenecks on deep, multi-hop tasks. This allows OmniSeek to decisively surpass OmniVideo-R1 on challenging datasets like VideoHolmes (74.6% vs. 62.9%) and OmniVideoBench (47.7% vs. 44.8%). Remarkably, OmniSeek's performance not only leads the open-source community but also remains competitive against closed-source models, outperforming Gemini-2.5-Pro on FutureOmni, VideoHolmes, and JointAV.

General Video Understanding. We also evaluate OmniSeek on general video understanding benchmarks to verify that our training paradigm preserves foundational perception capabilities. As in Table 3, OmniSeek not only retains the base model's strengths but actively improves upon them. On Video-MME and MLVU, OmniSeek achieves **78.5%** and **77.1%**, consistently outperforming the base Qwen3-Omni-Instruct. Despite being a 30B model, OmniSeek surpasses much larger models such as Qwen2.5-VL (72B) and LLaVA-OneVision (72B) across these tasks. Furthermore, compared to text-only reasoning approaches like OmniVideo-R1, OmniSeek maintains a clear advantage (e.g., 78.5% vs. 73.6% on Video-MME). These results indicate that the proposed training strategy preserves, and in several cases improves, the model's general video understanding capabilities despite being optimized for agentic audio-visual reasoning.

Number of Tool Calls. To further understand OmniSeek's active reasoning behavior, we analyze the correlation between the number of executed tool calls and the resulting accuracy in Figure 5. The preferred number of tool calls tends to reflect their reasoning complexity. For cross-modal correlation tasks like JointAVBench, accuracy peaks early at exactly 3 tool calls, increasing from 68.1% (2 calls) to 80.8%. Similarly, for Daily-Omni, accuracy increases to 87.1% at four tool calls. Conversely, on long-form reasoning benchmarks such as VideoHolmes and MMOU, OmniSeek benefits from more exploration, achieving peak performance (75.9% and 73.3%, respectively) when executing $\geq$5 tool calls. This suggests that OmniSeek varies its retrieval depth across tasks: it efficiently terminates early on bounded tasks while maintaining deep, long-horizon retrieval for buried clues when additional inspection is beneficial.

5.2. In-depth Analysis

Effectiveness of each Training Phase. Table 2 reveals three key training dynamics. First, we observe an "alignment tax" [10, 30] during Phase 1 SFT (Row b), where imposing a rigid

Model / Variant	Reasoning Paradigm	Tool Calling	Daily-Omni	WorldSense	FutureOmni	OmniVideoTest	OmniVideoBench	LVOmni
(a) Base Model	-	✗	71.9	55.1	53.6	54.5	43.6	35.8
(b) Text-only CoT	Single-turn text	✗	73.0	54.3	55.7	56.0	44.1	40.5
(c) OmniSeek (ours)	Multi-turn multi-modal	✓	80.0	62.4	58.3	69.5	47.7	44.2
Δ vs. Text-only CoT	-	-	+7.0	+8.1	+2.6	+13.5	+3.6	+3.7

Table 4. Ablation study comparing different reasoning paradigms.

Training Data	Training Format	Daily-Omni	WorldSense	FutureOmni	OmniVideoTest	OmniVideoBench	LVOmni	Video-MME
(a) Base Model	Zero-shot	71.9	55.1	53.6	54.5	43.6	35.8	76.8
(b) OmniVideo-100K	SFT + Single-turn QA	73.7	56.0	56.1	61.7	43.3	40.9	78.0
(c) OmniTraj-170K	SFT + Single-turn QA	75.8	56.0	57.1	59.2	45.0	41.6	77.6
Δ vs. Base Model	-	+3.9	+0.9	+3.5	+4.7	+1.4	+5.8	+0.8

Table 5. Analysis on the utility of the OmniTraj-170K corpus.

multi-turn tool-calling format temporarily disrupts the base model's pre-trained knowledge and generalization, leading to performance dips on most benchmarks like Daily-Omni (71.9% → 69.1%) and WorldSense (55.1% → 50.3%). However, Phase 2 RL (Row c) successfully recovers this degradation. By optimizing for outcome-based rewards rather than behavioral cloning, the agent learns strategic tool utilization. This triggers massive recoveries across the board, rapidly pushing LVOmni from 35.0% to 41.6% and driving a dramatic leap on VideoHolmes from 55.9% to 69.6%. Second, advancing to Phase 3 RL with a base rollout of $G = 8$ (Row d) further refines the policy, yielding substantial gains on complex reasoning datasets such as WorldSense (55.0% → 58.6%) and VideoHolmes (69.6% → 72.6%). Furthermore, scaling this GSPO rollout size from $G = 8$ (Row d) to $G = 16$ (Row e) provides consistent improvements. This indicates that broadening the search space is beneficial for exploration of alternative trajectories in complex audio-visual scenarios, increasing Daily-Omni from 75.9% to 78.2% and FutureOmni from 56.4% to 58.1%. Finally, integrating the Audio-Visual Necessity reward, r_{avn} (Row f), provides further gains. By explicitly penalizing unimodal shortcuts and forcing the agent to seek genuine cross-modal evidence, r_{avn} contributes additional gains, such as +3.2% on WorldSense and +2.2% on OmniVideoTest compared to the standard RL baseline (Row e). Combined, the complete pipeline achieves its best performance, yielding profound overall margins over the base model across diverse scenarios, including +8.1% on Daily-Omni, +7.3% on WorldSense, +4.1% on OmniVideoBench, and +8.4% on LVOmni.

Multi-turn tool-use vs. Single-turn textual reasoning. To isolate the impact of multi-turn tool-use, we compare OmniSeek against a text-only CoT baseline in Table 4. To ensure a fair comparison, this text-only CoT model undergoes the same two-stage RL training as OmniSeek. However, its action space is strictly constrained to text generation, forcing the model to rely entirely on prolonged internal textual thoughts to maximize outcome-based rewards without external tools. In Row (b), while this pure textual reasoning provides marginal gains on benchmarks like Daily-Omni (71.9% → 73.0%), it sometimes bottlenecks, or even degrades on tasks like WorldSense (55.1% → 54.3%). This highlights a limitation of text-only reasoning paradigms: prolonged textual thoughts cannot compensate for the missing context in long-form audio-video streams. Conversely, OmniSeek (Row c) strongly outperforms it. By actively interacting with the environment to additionally fetch targeted visual and audio clips, OmniSeek achieves substantial improvements over the Text-only CoT baseline. Notably, the multi-turn multimodal paradigm yields margins (Δ) of +13.5% on OmniVideoTest, +8.1% on WorldSense, and +7.0% on Daily-Omni. These results provide evidence that for complex omni reasoning, interleaved audio-video evidence provides substantial advantages over isolated textual scaling.

Utility of the OmniTraj-170K Corpus. To evaluate the utility of OmniTraj-170K as training data, we design experiments under a controlled single-turn SFT paradigm (Table 5), where we train the model to directly output the final answer. For a fair comparison against the OmniVideo-100K (Row b), we randomly sample 100K instances from our corpus using a 7:3 open-ended and multiple-choice mix. As shown in Row (c), training on this subset consistently outperforms the zero-shot Base Model (Row a), yielding robust gains on LVOmni (+5.8%) and Daily-Omni (+3.9%). Compared to OmniVideo-100K, our corpus demonstrates broader generalization, leading on Daily-Omni (+2.1%) and OmniVideoBench (+1.7%). Although OmniVideo-100K achieves a higher score on OmniVideoTest (61.7% vs. 59.2%), this advantage is expected because OmniVideo-100K and OmniVideoTest share the exact same origin, distribution, and stylistic design [2]. Crucially, OmniTraj-170K still achieves a solid +4.7% improvement over the base model on this benchmark, indicating its generalization without task-specific tuning.

6. Conclusion

We presented OmniSeek, an agentic framework for Omni-LLMs that turns audio-visual reasoning from passive perception into active, multi-turn evidence seeking. OmniSeek allows the model to adaptively retrieve cross-modal evidence

rather than relying on a single holistic encoding. We also construct OmniTraj-170K, a large-scale corpus of reasoning trajectories with interleaved modalities, together with a three phase training strategy and an Audio-Visual Necessity objective that discourages unimodal shortcuts. Extensive experiments show improvements on a wide range of audio-visual benchmarks requiring long-form and multi-hop reasoning.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[2] Xinyue Cai, Chaoyou Fu, Yi-Fan Zhang, Ran He, and Caifeng Shan. Omnivideo-100k: A dataset for audio-visual reasoning through structured scripts and evidence chains. *arXiv preprint arXiv:2606.14702*, 2026.

[3] Brandon Castellano. PySceneDetect: Python and OpenCV-based scene cut/transition detection, 2014.

[4] Jianghan Chao, Wenhui Tan, Yuchong Sun, Ruihua Song, Liyun Ru, et al. Jointavbench: A benchmark for joint audio-visual reasoning evaluation. In *International Conference on Learning Representations*, pages 84035–84062, 2026.

[5] Qian Chen, Jinlan Fu, Changsong Li, Min Zhang, See-Kiong Ng, and Xipeng Qiu. Futureomni: Evaluating future forecasting from omni-modal context for multimodal llms. In *International conference on machine learning*, 2026.

[6] Yu Chen, Caorui Li, Ziyu Xiong, Yidong Wang, Mingqi Gao, Shuman Liu, Biao Liu, Chunfeng Yang, Anxiang Zeng, Haibo Zhang, et al. Omnireasoner: Thinking with long audio-video via native tool use. *arXiv preprint arXiv:2607.19339*, 2026.

[7] Zhangquan Chen, Jiale Tao, Ruihuang Li, Yihao Hu, Ruitao Chen, Zhantao Yang, Xinlei Yu, Haodong Jing, Manyuan Zhang, Shuai Shao, et al. Omnivideo-r1: Reinforcing audio-visual reasoning with query intention and modality attention. In *International conference on machine learning*, 2026.

[8] Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? In *European Conference on Computer Vision*, 2026.

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

[10] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. In *International conference on machine learning*, 2025.

[11] 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.

[12] Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

[13] Yifan Dai, Zhenhua Wu, Bohan Zeng, Daili Hua, Jialing Liu, Bozhou Li, Yuran Wang, Chengzhuo Tong, Hao Liang, Xiaochen Ma, et al. Latentomni: Rethinking omni-modal understanding via unified audio-visual latent reasoning. *arXiv preprint arXiv:2605.22012*, 2026.

[14] Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, Matthieu Le, Tyler Poon, Danial Mohseni Taheri, Ilia Karmanov, Guilin Liu, Jarno Seppanen, Arushi Goel, et al. Nemotron 3 nano omni: efficient and open multimodal intelligence. *arXiv preprint arXiv:2604.24954*, 2026.

[15] Yue Ding, Yiyan Ji, Jungang Li, Xuyang Liu, Xinlong Chen, Junfei Wu, Bozhou Li, Bohan Zeng, Yang Shi, Yushuo Guan, et al. Omnisift: Modality-asymmetric token compression for efficient omni-modal large language models. *arXiv preprint arXiv:2602.04804*, 2026.

[16] Yang Ding, Xin Lai, Yizhen Zhang, Wei Li, Ruihang Chu, and Yujiu Yang. Videozoomer: Reinforcement-learned temporal focusing for long video reasoning. In *International Conference on Learning Representations*, pages 20087–20111, 2026.

[17] Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Xinze Guan, and Xin Wang. Grit: Teaching mllms to think with images. *Advances in Neural Information Processing Systems*, 38:116522–116543, 2026.

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

[19] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2026.

[20] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Yuhang Dai, Meng Zhao, Yi-Fan Zhang, Shaoqi Dong, Yangze Li, Xiong Wang, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024.

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

[22] Chaoyou Fu, Haojia Lin, Xiong Wang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long MA, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *Advances in Neural Information Processing Systems*, 38:75300–75320, 2026.

[23] Mingfei Gao, Rui Tian, Haiming Gang, Bohan Zhai, Le Zhang, Yuanzheng Gong, Di Feng, Ege Özsoy, Kaixin Ma, Vishwesh Kirthivasan, Oğuzhan Fatih Kar, Roman Bachmann, Anders Boesen Lindbo Larsen, and Afshin Dehghan. Mintact: A unified visual agent for digital environments. *arXiv preprint arXiv:2609.22083*, 2026.

[24] Arushi Goel, Sreyan Ghosh, Vatsal Agarwal, Nishit Anand, Kaousheik Jayakumar, Lasha Koroshinadze, Yao Xu, Katie Lyons, James Case, Karan Sapra, et al. Mmou: A massive multi-task omni understanding and reasoning benchmark for long and complex real-world videos. *arXiv preprint arXiv:2603.14145*, 2026.

[25] Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. In *International Conference on Learning Representations*, pages 52423–52443, 2026.

[26] Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, and Guohai Xu. Deepeyesv2: Toward agentic multimodal model. In *International Conference on Learning Representations*, pages 114851–114872, 2026.

[27] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14271–14280. IEEE, 2024.

[28] Jitesh Jain, Zhengyuan Yang, Humphrey Shi, Jianfeng Gao, and Jianwei Yang. Ola-vlm: Elevating visual perception in multimodal llms with auxiliary embedding distillation. *arXiv preprint arXiv:2412.09585*, 2024.

[29] Chaoya Jiang, Yongrui Heng, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. Vlm-r^3: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. *Advances in Neural Information Processing Systems*, 38:63841–63869, 2026.

[30] Feiyang Kang, Michael Kuchnik, Karthik Padthe, Marin Vlastelica, Ruoxi Jia, Carole-Jean Wu, and Newsha Ardalani. Quagmires in sft-rl post-training: When high sft scores mislead and what to use instead. In *International Conference on Learning Representations*, pages 56876–56918, 2026.

[31] Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. In *International Conference on Learning Representations*, pages 76722–76746, 2026.

[32] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[33] Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Zhenghao Song, Dingling Zhang, Ying He, et al. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. In *International Conference on Learning Representations*, pages 138214–138236, 2026.

[34] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

[35] Yunxin Li, Xinyu Chen, Shenyuan Jiang, Haoyuan Shi, Zhenyu Liu, Xuanyu Zhang, Nanhao Deng, Zhenran Xu, Yicheng Ma, Meishan Zhang, et al. Uni-moe-2.0-omni: Scaling language-centric omnimodal large model with advanced moe, training and data. *arXiv preprint arXiv:2511.12609*, 2025.

[36] 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*, 2025.

[37] Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, et al. Open-o3-video: Grounded video reasoning with explicit spatio-temporal evidence. In *International conference on machine learning*, 2026.

[38] Qwen Team. Qwen3.5: Towards native multimodal agents, 2026.

[39] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14313–14323. IEEE, 2024.

[40] Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.

[41] Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. *Advances in Neural Information Processing Systems*, 38:8222–8251, 2026.

[42] Wanshun Su, Yang Shi, Feihu Liu, Ziwen Yu, Yan Min, Zhuoran Zhang, Qixun Wang, Haotian Wang, Shixuan Liu, Yuanxing Zhang, et al. Omnipack: Unified token compression for efficient omni-modal large language models. *arXiv preprint arXiv:2608.03812*, 2026.

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

[44] Keda Tao, Kele Shao, Bohan Yu, Weiqiang Wang, Jian Liu, and Huan Wang. Omnizip: Audio-guided dynamic token compression for fast omnimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17682–17692, 2026.

[45] 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.

[46] 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.

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

[48] Omkar Thawakar, Dinura Dissanayake, Ketan Pravin More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Ilmuz Zaman Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24290–24315, 2025.

[49] Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. *arXiv preprint arXiv:2410.03290*, 2024.

[50] Haibo Wang, Bo Feng, Zhengfeng Lai, Mingze Xu, Shiyu Li, Weifeng Ge, Afshin Dehghan, Meng Cao, and Ping Huang. Streambridge: Turning your offline video large language model into a proactive streaming assistant. *Advances in Neural Information Processing Systems*, 38:132332–132359, 2026.

[51] 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*, 2025.

[52] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22958–22967. IEEE, 2025.

[53] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

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

[55] Zhenghao Xing, Ruiyang Xu, Yuxuan Wang, Jinzheng He, Ziyang Ma, Qize Yang, Yunfei Chu, Jin Xu, Junyang Lin, Chi-Wing Fu, et al. Native active perception as reasoning for omni-modal understanding. In *International conference on machine learning*, 2026.

[56] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Ke qin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *ArXiv*, abs/2503.20215, 2025.

[57] 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*, 2025.

[58] Mingze Xu, Mingfei Gao, Shiyu Li, Jiasen Lu, Zhe Gan, Zhengfeng Lai, Meng Cao, Kai Kang, Yinfei Yang, and Afshin Dehghan. Slowfast-llava-1.5: A family of token-efficient video large language models for long-form video understanding. *Conference on Language Modeling*, 2025.

[59] Ziang Yan, Yinan He, Xinhao Li, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *Advances in Neural Information Processing Systems*, 38:119152–119184, 2026.

[60] 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 *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6580–6598, 2025.

[61] Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, et al. Longvt: Incentivizing" thinking with long videos" via native tool calling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 33816–33826, 2026.

[62] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

[63] Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Zhen Wan, Jinchuan Tian, An-Chieh Cheng, Ligeng Zhu, Yuanhang Su, Yuming Lou, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. In *International Conference on Learning Representations*, pages 56101–56138, 2026.

[64] Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, et al. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. In *International conference on machine learning*, 2026.

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

[66] Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 32903–32914, 2026.

[67] Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379*, 2025.

[68] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization, 2025.

[69] Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, and Chao Shen. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, pages 126775–126798, 2026.

[70] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *2025 IEEE/CVF Conference on Computer*

Vision and Pattern Recognition (CVPR), pages 13691–13701. IEEE, 2025.

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

A. Appendix

A.1. More Implementation details

We show the details of each training phase in Table 6.

A.2. Benchmarks

Daily-Omni [71]: a multiple-choice audio–visual QA benchmark targeting cross-modal *temporal* reasoning over everyday scenes. It draws 684 real-world videos segmented into 30- and 60-second clips to probe different temporal contexts, and pairs them with 1,197 questions organized into six task families, each constructed so that the answer requires temporally aligning auditory and visual cues.

AVUT [60]: an audio-centric video understanding benchmark that evaluates comprehension with an explicit focus on auditory information and targets the "text-shortcut" problem prevalent in prior benchmarks. We use its expert-annotated subset (AV-Human), comprising 1,734 human-written QA pairs over 698 YouTube videos from audio-centric domains.

WorldSense [25]: a benchmark for real-world omni-modal understanding whose tasks are designed around a tight coupling of audio and video, so that neither stream alone suffices. It collects 1,662 audio-visually synchronized videos (average length 141s) organized into 8 primary domains and 67 fine-grained subcategories, and provides 3,172 multiple-choice questions spanning 26 cognitive tasks that range from low-level perception to high-level reasoning.

FutureOmni [5]: the first benchmark designed to evaluate omni-modal *future forecasting* from audio–visual environments, requiring cross-modal causal and temporal reasoning together with internal knowledge to anticipate events yet to occur. It contains 919 videos and 1,034 multiple-choice QA pairs across 8 primary domains.

OmniVideo-Test [2]: the human-verified test split of the OmniVideo-100K, constructed via entity-anchored video scripting and clue-guided QA generation so that questions carry long-term temporal spans and cross-modal dependencies. It comprises 505 human-verified multiple-choice QA pairs over 264 videos, spanning 10 audio-visual task types.

VideoHolmes [8]: a benchmark for "Holmes-style" complex video reasoning, in which a model must actively locate and connect visual clues scattered across a video to infer the answer. It is built from 270 manually annotated suspense short films (1–5 minutes, sourced from YouTube) and comprises 1,837 questions across seven reasoning tasks centered on causal and multi-clue inference. Notably, all training samples used in our Phase 2 training are strictly disjoint from the evaluation sets at both the question and video levels. In particular, the 1K VideoHolmes training examples are drawn exclusively from a non-overlapping training split, with no videos or questions shared with the reported VideoHolmes evaluation set. The same strict separation is applied between OmniVideo100K training samples and OmniVideo-Test.

JointAVBench [4]: a benchmark with strict audio–video correlation, designed so that questions cannot be answered from a single modality. It consists of 2,853 manually verified multiple-choice questions built from professionally produced films, spanning five cognitive dimensions, four audio information types (speech, sound events, music, and vocal traits), and three scene spans (single-, cross-, and full-scene).

OmniVideoBench [33]: a large-scale benchmark for synergistic audio–visual reasoning that stresses modality complementarity and logical consistency. It contains 1,000 manually verified question–answer pairs, each accompanied by a step-by-step reasoning trace, drawn from 628 videos (several seconds to 30 minutes) that span 8 major categories and 68 subcategories of real-world content such as news, sports, documentaries, and vlogs.

MMOU [24]: a massive multi-task omni understanding and reasoning benchmark for long and complex real-world videos, designed to test joint reasoning over visual, audio, and textual signals rather than any modality in isolation. It contains 20,000 expertly annotated multiple-choice questions over 11,877 long-form web videos (average duration $\sim$522s), organized into 13 fundamental audio-visual skill categories (with an average of three skills per question) and spanning 10 major domains and 35 fine-grained subcategories; each question is posed against 10 answer options (one correct and nine hard distractors), making the benchmark notably challenging. The 20,000 questions are divided into a *test* split (15K) and a *test-mini* split (5K); we evaluate on the test-mini split. Results denoted with † in Table 1 are evaluated on the test split.

LVOmniBench [45]: a benchmark dedicated to *long-form* audio–video understanding for Omni-LLMs, addressing the gap that prior evaluations focus on clips under five minutes. It comprises 275 manually selected and annotated videos ranging from 10 to 90 minutes and 1,014 QA pairs, evaluating long-term memory, temporal localization, fine-grained understanding, and multimodal perception. Due to the extreme duration of certain videos causing the raw audio tokens to exceed the model's maximum sequence length, we apply speedup to videos exceeding 30 minutes to cap their effective duration within 30 minutes.

Video-MME [21]: the first comprehensive evaluation benchmark for multi-modal LLMs in video analysis, spanning a wide range of visual domains and video lengths. It curates 900 videos (11 seconds to 1 hour) across six primary domains and annotates 2,700 high-quality multiple-choice questions (three per video), reported over short, medium, and long duration splits. For these general video benchmarks, we follow their conventional evaluation protocols and use a single-turn direct-answer setting, without tool invocation or multi-turn interaction.

MLVU [70]: a multi-task benchmark for long-video understanding built from videos of diversified lengths ranging

	Phase 1 (SFT)	Phase 2 (GSPO)	Phase 3 (GSPO)
Training Data & Hardware			
Dataset	OmniTraj-170K	OmniVideo100K + VideoHolmes	OmniTraj-170K
Number of Samples	170K (90% QA + 10% Tool)	30K (only Multi-Choice) + 1K	8K (Hard)
Epochs	1	1	1
Train Batch Size	128	256	256
Number of GPUs	32 × H200	128 × H200	128 × H200
Optimization			
Optimizer	AdamW	AdamW	AdamW
Learning Rate	5×10^{-6}	1×10^{-6}	1×10^{-6}
LR Scheduler	cosine	cosine	cosine
Warmup Ratio	0.03	0.1	0.1
KL Coefficient (β)	-	0	0
Gradient Updates per Batch	-	4	4
Clip Ratio (low / high)	-	$3e^{-4}$/ $4e^{-4}$	$3e^{-4}$/ $4e^{-4}$
Reward	-	$r_{\text{acc}} + r_{\text{format}} + r_{\text{tool}}$	$r_{\text{acc}} + r_{\text{format}} + r_{\text{tool}} + r_{\text{avn}}$
Environment & Rollout			
Rollout Size (G)	-	8	16
Temperature/Top-p/Top-K	-	1/1/-1	1/1/-1
Max Interaction Turns	-	8	8
Max Sequence Length	32768	65536	65536
Model			
Freeze Vision/Audio Encoder	True	True	True
FPS	2	2	2
MAX_FRAME	256	128	128
MAX_PIXELS	$256 \times 32 \times 32$	$128 \times 32 \times 32$	$128 \times 32 \times 32$
CLIP_MAX_FRAME	45	45	45
CLIP_MAX_PIXELS	$256 \times 32 \times 32$	$256 \times 32 \times 32$	$256 \times 32 \times 32$
CLIP_MAX_AUDIO_SECONDS	60	60	60

Table 6. Implementation and Hyperparameter Details for the Three-Phase Training Strategy.

from 3 minutes to 2 hours (averaging ∼15 minutes) and spanning diverse genres such as movies, egocentric footage, documentaries, and surveillance. It comprises 3,102 questions across nine distinct tasks (a development set of 2,593 and a test set of 509), covering both close-ended and open-ended formats.

LongVideoBench [54]: a benchmark for long-context video–language understanding, centered on *referring reasoning* questions that require retrieving and reasoning over specific referred moments in long multimodal inputs. It contains 3,763 web-collected videos with subtitles (up to one hour long, across themes such as daily life, movies, knowledge, and news) and 6,678 human-crafted multiple-choice questions organized into 17 fine-grained categories.

LVBench [52]: an extreme long-video understanding benchmark targeting videos far longer than prior datasets, defining long videos as those lasting at least 30 minutes. It consists of 103 manually filtered high-quality videos totaling 117 hours (averaging ∼4,101 seconds, i.e. ∼68 minutes) and 1,549 question–answer pairs spanning multiple task categories that demand long-term temporal understanding.

A.3. Visual Annotation Prompt

To ensure strict modality isolation and prevent hallucination during the annotation phase, we use a highly constrained system prompt. The prompt explicitly forbids the use of auditory cues and enforces the rejection of segments containing overlaid dialogue subtitles, guaranteeing that the generated visual events are derived entirely from visual evidence. The complete system prompt is provided in 1.

Listing 1. System prompt for generating dense, timestamped visual captions per scene.

```
You are an expert visual annotator for audio-visual
    temporal grounding data generation.

You will be given a short video segment. Your task is
    to generate a precise visual-only annotation
    that can later be used to create cross-modal QA
    pairs with ASR.

Focus only on what is visible in the video segment.

Do not use audio, dialogue, ASR, subtitles inferred
    from speech, prior context, outside knowledge,
    or common-sense story assumptions to infer
    visual content.
```

```
Output valid JSON only.

Visual annotation requirements:

1. Describe concrete visible evidence, including
     people, actions, objects, scene/location, camera
      focus, shot transitions, visible on-screen text
     , screens, signs, labels, cards, maps, documents
     , or product packaging.
2. Preserve visible text exactly when readable.
3. If visible text is partially readable, mark it as
     partially readable instead of guessing.
4. Do not hallucinate names, identities,
     relationships, emotions, intentions, brands,
     locations, or events.
5. If the role or identity is uncertain, describe
     visible attributes instead of guessing.
6. If an object category is uncertain, use
     conservative wording such as "appears to be" or
     "possibly".
7. Describe visual events in chronological order.
8. Assign each visual event an approximate timestamp
     range relative to the start of this segment.
9. Prefer specific, localizable visual details over
     generic scene summaries.
10. Avoid vague descriptions such as "someone is
     doing something", "a scene is shown", or "
     various objects are visible".
11. Do not include audio content or what anyone says.
12. Do not say "the video shows", "this clip shows",
     "we see", or other meta-video phrases inside the
      annotation fields.
13. Keep the annotation concise but information-dense
     .

Timestamp rules:

* All visual event timestamps must be relative to the
      start of the current segment.
* Use seconds as numbers, e.g., [0.0, 2.3].
* The timestamp range should cover when the visual
     evidence is visible or happening.
* If the event is visible throughout the whole
     segment, use [0.0, segment_duration].
* If exact timing is uncertain, give a conservative
     approximate range.
* Do not create very narrow timestamps unless the
     visual event is clearly brief.
* visual_events must be sorted by timestamp start
     time.

Output schema:
{
  "visual_events": [
    {
      "timestamp": [number, number],
      "description": string
    }
  ]
}

Field definitions:

* "visual_events": Chronological visible events or
     states. Each item should be concrete and
     visually grounded.
* "visual_events.timestamp": An [start, end] range in
      seconds, relative to the start of this segment.
* "visual_events.description": A concrete description
      of the visible event, object, action, text,
     screen, sign, card, or camera focus during that
     timestamp range.

Important:

* If the segment has no distinctive visual evidence,
     return:
  {
    "visual_events": []
  }
* Do not output markdown.
* Do not output explanations.
* Do not output extra fields.
* If the segment contains bottom subtitles, closed
     captions, dialogue captions, lyric captions, or
     any speech-transcription text overlaid near the
     bottom of the frame, reject the entire segment
     and return:
  {
    "visual_events": []
  }
* Do not annotate bottom subtitles as visible text.

JSON formatting requirements:
- Return valid JSON only.
- Escape all double quotes inside string values using
      backslashes.
- For visible text inside descriptions, use single
     quotes instead of double quotes when possible.
- The first character must be "{".
- The last character must be "}".
- Do not use markdown code fences.
- Do not output any text before or after the JSON.
```

A.4. Evidence-Grounded QA Generation Prompt

In the second stage of our data engine, we synthesize multi-hop audio-visual questions based on the structured context derived from the first stage. To ensure that the questions demand cross-modal reasoning and cannot be answered via single-modality shortcuts or language priors, we employ a highly detailed system prompt. This prompt not only defines 19 diverse question types but also enforces the explicit planning of an *evidence chain* with deterministic timestamps and content directly copied from the Stage 1 annotations. The complete system prompt, excluding the full list of the 19 question types for brevity, is provided in Listing 2.

Listing 2. System prompt for generating Evidence-Grounded QA and cross-modal evidence chains.

```
You are an expert in audio-visual video understanding
      and multiple-choice QA generation.

You will receive the annotation for ONE video:
* "overall_description": a human-readable description
      of the WHOLE video, for reference/consistency.
* "scenes": a list of segments (each scene is one
     segment). Reason over the WHOLE video, not a
     single scene: a QA may draw evidence from any
     scene(s), and cross-scene QA connecting events
     across different segments are encouraged. Each
     scene contains:
  - segment_id, start, end, duration
  - caption.visual_events: [{ "timestamp": [start,
      end], "description": <visual text> }]
  - asr: [{ "start", "end", "text": <spoken text> }]

Definitions:
* video evidence = a caption.visual_events.
     description (with its timestamp)
* audio evidence = an asr utterance text (with its
     start/end)
* Every QA is GROUNDED in specific evidence spans
     that you must list explicitly.

### Consistency (use "overall_description")
Segments were captioned INDEPENDENTLY - the annotator
      of one segment could not see the others - so
     the SAME person/object may be described
     differently across segments (e.g. "a man in a
```

```
    red shirt" in one segment and "the host" in
    another may be the same person). Use "
    overall_description" as the ground truth about
    the whole video to reconcile these: decide when
    differently-worded segment captions refer to the
     same entity, and phrase your question, options,
     and answer with a CONSISTENT reference for that
     entity. Never contradict "overall_description".
     However, the evidence spans "text" must still
    be copied from the segment visual_events/asr (
    not from overall_description);
    overall_description guides your wording and
    reasoning, it is not itself an evidence source.

### Goal
Generate multiple-choice QA that HELP a human
    understand the video and that can be answered
    with and ONLY with BOTH audio (speech/sound) AND
     visual information - never answerable from one
    modality alone, from the question wording, or
    from outside knowledge / language priors.

### Question Types (generate a diverse mix FROM the
    NINETEEN types below; put the exact label in "
    type")
Do NOT force every type to appear in one video.
    Select only types that are strongly supported by
     clean, unambiguous evidence. Each type below
    states its cross-modal binding -- the
    construction that forces both modalities.
- "AV Event Alignment": given an event in one
    modality, identify the event in the OTHER
    modality that occurs simultaneously with it.
    Binding: anchor event in modality X at time t ->
     answer is the co-occurring event in modality Y
    at the same t.
- "Scene Transformation Detection": use an AUDIO cue
    to pinpoint a moment of VISUAL change, and
    identify how the scene shifts (from -> to).
    Binding: audio marks when; the answer is the
    visual from->to change.
...
(17 additional question types are detailed here: "
    Context Understanding", "Comparative", "Event
    Sequence", "Reordering", "Causal Reasoning", "
    Inference", "Fine-grained Perception", "Spatial
    Reasoning", "Sentiment Analysis", "Reference
    Reasoning", "Relationship Reasoning", "
    Summarization", "Speaker Grounding", "Visual
    Text & Diagram Grounding", "Cross-modal
    Consistency Checking", "Counting", and "Object
    Interaction & State Change".)
...

### Multiple-choice rules
- Exactly 4 options. Options are PLAIN TEXT with NO
    letter/number prefix (no "A.", no "1)").
- Exactly one option is correct; the other three are
    wrong but plausible.
- Make distractors deceptive: similar in content/
    length to the answer so the question needs
    careful attention to both modalities; avoid one
    obviously-odd option. Keep all four options
    similar length to avoid length bias.
- Do not leak the answer in the question; the
    question must not contain or closely paraphrase
    the correct option, and must not state so much
    that it is answerable without the video.
- Set "answer_index" to the 0-based position of the
    correct option, "answer" to that options text,
    and "answer_label" to its letter (A/B/C/D). (
    Correct-option position may be re-randomized
    later.)

### Requirements
- REQUIRE BOTH MODALITIES: answering must need at
    least one audio AND at least one visual evidence
     span. Reject any QA answerable from audio alone
    , video alone, or the question text alone.
- Ground strictly in the provided visual_events and
    asr. Do NOT use outside knowledge; do NOT invent
     or infer objects, actions, names, numbers,
    intentions, emotions, or timestamps beyond the
    evidence.
- Prefer precision over quantity. Generate AT MOST 5
    QA for the WHOLE video (not per scene); across
    those, favor a diverse mix of the nineteen types
     and include cross-scene QA where the video
    supports them. If no reliable AV QA can be made,
     return [].
- Reject generic/repeated visuals (blank/solid-color
    screens, static logos, repeated talking-head or
    text shots) and vague references ("the person/
    scene/screen/speaker") unless the evidence is
    distinctive enough to identify one unique span.
    The question must not mention timestamps and
    must not be yes/no.

### The "evidence" field (REQUIRED for every QA)
List the exact input spans the QA is built on.
    MODALITY, ORDER and COUNT match the QA.
- Use BETWEEN 2 AND 7 evidence spans (inclusive).
- Order evidence in the LOGICAL order the reasoning
    visits it - i.e. which span you would look at or
     listen to FIRST to start answering, then next,
    and so on until the answer. This is the
    retrieval order a solver would follow (e.g.
    locate the clue span before the answer-bearing
    span); it is NOT necessarily chronological by
    timestamp.
- Each span: { "modality": "audio"|"video", "
    timestamp": [start,end], "text": <copied text>,
    "answer_depends_on": true|false }
- answer_depends_on=true -> this span directly
    determines the correct answer (target).
  answer_depends_on=false -> this span is context/
     clue used to frame or locate the question.
- At least ONE span must be answer_depends_on=true,
    and the evidence overall MUST include at least
    one "audio" span and at least one "video" span.
- SINGLE-EVENT, NO AMBIGUITY (the binding rule): each
     spans "text" must precisely and unambiguously
    describe exactly ONE event happening in that
    time window - nothing else. Because segment
    captions can be coarse, a visual_event span may
    actually contain SEVERAL distinct visual events
    lumped together. If your question concerns only
    one of those events but the span (and its text)
    also covers other events, that span is AMBIGUOUS
     (multiple video events map onto one audio event
    , or vice versa) - in that case DO NOT create
    the QA. Only build a QA when every evidence span
     cleanly isolates the one event the QA refers to.
     If no span cleanly isolates the needed event,
    skip the QA.
- SPAN LENGTH: keep each span roughly BETWEEN 3 AND
    20 SECONDS - long enough to cover one whole
    event, short enough not to lump in others. This
    range is a guideline serving the binding SINGLE-
    EVENT rule above (a shorter clean utterance that
     already isolates one event is acceptable); what
     matters is that the span isolates exactly one
    unambiguous event.
- Guide (not a hard rule): follow the chosen types "
    Binding" clause above to pick spans - include
    the clue/anchor span(s) that locate or frame the
     question (answer_depends_on=false) plus the
    answer-bearing span(s) the answer is read from (
    answer_depends_on=true), one span per distinct
    event the QA truly needs, always covering both
    modalities.

### Evidence text rules
- Audio text: copy the asr text EXACTLY (do not
    correct, normalize, translate, or rewrite).
- Video text: use visual_events.description; you may
    drop irrelevant clauses but do not paraphrase or
     add details. Never mix audio content into a
```

```
    video span or vice versa.

### Timestamp rules
- Audio spans use exact asr start/end; video spans
    use exact visual_events.timestamp. Do NOT
    substitute segment start/end for a visual events
     timestamp. Reject spans with missing/invalid
    timestamps or outside their segment boundary.
- Keep each evidence span roughly 3-20 seconds (
    guideline for the single-event rule above).

### Language rules
- Only use English asr; ignore non-English asr (do
    not build QA depending on it). Write every
    question, option and answer in English.

### Output
Return valid JSON only: a JSON list, no markdown, no
    comments, no extra top-level fields.
If nothing reliable can be generated, return [].
Each item:
[
  {
    "question": string,
    "type": "AV Event Alignment" | "Scene
        Transformation Detection" | ... | "Object
        Interaction & State Change",
    "options": [string, string, string, string],
    "answer": string,
    "answer_index": integer,
    "answer_label": "A" | "B" | "C" | "D",
    "evidence": [
      { "modality": "audio" | "video", "timestamp": [
          number, number], "text": string, "
          answer_depends_on": false },
      { "modality": "audio" | "video", "timestamp": [
          number, number], "text": string, "
          answer_depends_on": true }
    ]
  }
]

### Final validation (silently drop any item that
    fails)
- Does answering truly need BOTH an audio and a
    visual evidence span?
- Unanswerable from question wording / one modality /
     outside knowledge alone?
- Exactly 4 plain-text options, one correct,
    distractors plausible & similar length, no
    answer leak?
- answer/answer_index/answer_label mutually
    consistent?
- Every evidence span grounded, unique, non-generic,
    valid timestamp, text copied per rules?
- Between 2 and 7 evidence spans?
- Does every evidence span cleanly isolate exactly
    ONE event (no ambiguity, nothing extra)? If any
    span lumps in other events that the QA does not
    concern, DROP the QA.
- >=1 answer_depends_on=true, and >=1 audio and >=1
    video span?
- Entity references consistent with
    overall_description (same entity referred to
    consistently)?
```

A.5. Interleaved Trajectory Assembly Prompt

In the final stage of our data engine, we use a specialized system prompt to instruct the model to generate the internal reasoning steps (<think>) and determine the logical retrieval order of the predefined evidence spans. Crucially, to prevent tool-use hallucinations, the model is restricted from generating the tool calls or observations itself; these are deterministically interleaved by our pipeline based on the model's chosen order. The prompt enforces strict rules against meta-commentary, premature answer leaking, and the exposure of internal bookkeeping indices. The system prompt for trajectory assembly is provided in Listing 3.

Listing 3. System prompt for generating internal reasoning steps and logical retrieval orders for interleaved trajectories.

```
You are writing the internal reasoning for a gold
    demonstration of how an omni model solves a
    multi-hop, cross-modal video question by
    inspecting video and audio clips with two tools:
     get_video_clip(start,end) and get_audio_clip(
    start,end).

You are given:
* optionally, the VIDEO DURATION and a VIDEO OVERALL
    DESCRIPTION of the whole video. These are
    background context ONLY - to help you phrase the
     reasoning naturally and consistently. They are
    NOT retrievable clips and must NOT be cited or
    treated as evidence spans; never say you looked
    at "the overall description".
* a QUESTION,
* the FINAL ANSWER (already correct - do not change
    it),
* a set of AVAILABLE EVIDENCE SPANS, each numbered [i
    ]. Each span is one clip you can retrieve, with
    its modality, time window, and the exact content
     it returns.
  IMPORTANT: these spans are listed in an ARBITRARY
     order (often just chronological by timestamp).
      That is NOT necessarily the order in which a
     solver should inspect them.

Your job:
1. Decide the LOGICAL retrieval order - which span
    you would inspect FIRST to start answering, then
     next, and so on - reasoning from the QUESTION
    and ANSWER (e.g. locate a clue span before the
    span that carries the answer). Output it as "
    order": a permutation of ALL the given span
    indices, in that logical order.
2. Write the model THINK steps and a short REASONING
    summary, consistent with that order.

You MUST use every span exactly once ("order" is a
    permutation of all indices). Do not add, drop,
    merge, or invent spans; the clip contents are
    fixed exactly as given.

The [i] numbers and the "order" list are INTERNAL
    bookkeeping only. The think strings must be
    natural first-person reasoning that NEVER
    mentions them. Refer to each clip naturally - by
     its modality and what you look at / listen for
    or what it shows / says (e.g. "let me look at
    the video near the start", "I'll listen to what
    the speaker says next", "the screen shows ...").

Produce exactly (N + 1) think strings, aligned to
    YOUR chosen order (N = number of spans):
* think[0]: the opening plan, BEFORE any clip is
    retrieved. Explain what the question requires
    and which clip you will inspect FIRST and why,
    referring to it naturally (modality + what you
    expect to find), NOT by an index. Do NOT state
    any clip content yet, and do NOT preview, guess,
     or hint at later findings or the answer.
* think[k] for k = 1 .. N-1 (after the k-th retrieved
     clip in YOUR order): FIRST restate, in your own
     words, the key content that clip just returned
    (for audio: "she says / mentions ..."; for video
    : "the screen shows / the background is ..."),
    THEN say what you will retrieve next and why (e.
    g. switching modality, moving to another moment)
    , naturally.
```

```
* think[N] (after the last clip): FIRST restate the
     last clip content, THEN connect the gathered
     evidence to what the question asks, so the
     answer follows naturally. Do not literally write
      the final answer sentence here; lead up to it.

Hard rules:
* NEVER mention span indices or bookkeeping in the
     thinks: no "span [0]", "span 3", "the first/
     final span", "index", "timestamp", "earliest/
     latest timestamp", or similar. Refer to clips
     only by modality and content. (Indices belong
     ONLY in the "order" field.)
* Do not justify a choice by timestamp order (e.g. "
     since it has the earliest timestamp"); justify
     it by the reasoning logic (what you need to find
      and why you look there).
* Use ONLY the information contained in the given
     clip contents. Never invent objects, names,
     actions, or details that are not in the clips.
* Every non-first think must clearly reflect the
     content of the span retrieved just before it (in
      your chosen order). Concatenated in order, the
     thinks must read as ONE logically complete,
     correct, fluent chain of reasoning that on its
     own leads to the answer. Use natural connectives
      (so, now, next, here, finally).
* Write like a person reasoning, grounded in the
     concrete evidence. Do NOT use formulaic meta-
     commentary such as "the evidence is sufficient",
      "sufficient to answer", "to address the query/
     question", or similar filler. End on the
     substance of the evidence.
* Keep each think concise (1-3 sentences). English
     only.
* The reasoning field is one or two sentences
     summarizing how the answer follows from the
     evidence.

Output valid JSON only, exactly this schema and
     nothing else:
{
  "reasoning": "one or two sentences",
  "order": [<span indices: a permutation of 0..N-1 in
       logical retrieval order>],
  "thinks": ["think[0]", "think[1]", ..., "think[N]"]
}
"order" length must be exactly N; "thinks" length
     must be exactly N+1.
Do not output markdown, code fences, comments, or any
      extra text.
```

A.6. Per-Benchmark Fine-Grained Results

We report the fine-grained breakdown results of OmniSeek on these benchmarks from Table 7 to Table 16, following the official dimensions defined by the respective benchmark.

A.7. Limitations, Discussions and Future Work

While OmniSeek demonstrates strong multi-turn reasoning capabilities, it encounters a structural bottleneck when processing extremely long-form videos (e.g., 1 to 2 hours), primarily due to the context limit imposed by continuous audio streams. Unlike visual inputs, which can be easily constrained by downsampling to a maximum number of frames (e.g., 128 frames), raw audio encoding scales linearly with time. For our base model, Qwen3-Omni-30B-A3B-Instruct, the native context length is bounded at 32K tokens. Given an audio tokenization rate of approximately 12.5 tokens per second (at a 16kHz sampling rate), two hours of audio generates roughly 90K audio tokens, immediately exceeding the model's maximum context capacity.

When this context window is breached, we observe a degradation in the model's instruction-following and formatting capabilities. For instance, in failure cases sampled from extremely long videos in LVOmniBench (e.g., an 88-minute video), the model's structured reasoning pattern collapses. As the context overflows, the agent loses the ability to invoke tools. Instead, it falls into a repetitive `<think>` loop, hallucinating sensory evidence directly within its internal reasoning blocks (e.g., fabricating observations like *"The audio reveals a voice..."* or *"The visual shows a woman in a gym setting..."*) without ever executing the `<tool_call>` or arriving at a valid `<answer>`.

To mitigate this in the current framework, we apply a linear playback speed-up to videos exceeding 30 minutes, effectively compressing the sequence to increase the information density per audio token and safely cap the length within the 32K window. However, this heuristic may inevitably distort fine-grained auditory cues such as speech pitch or environmental sound textures. In future work, we plan to address this limitation through two primary avenues. First, from the training perspective, we aim to natively expand the model's context window by incorporating longer multi-turn trajectories into the reinforcement learning pipeline. Second, from the inference perspective, we will explore modality-asymmetric token compression strategies, such as dynamically discarding silent audio tokens or merging redundant acoustic features to enable the agent to efficiently process hours-long multimodal streams without structural collapse.

	AV Align	Comp.	Ctx. Und.	Evt. Seq.	Infer.	Reas.	30s	60s	**Overall**
OmniSeek	73.9	80.2	77.7	76.8	89.6	88.0	78.5	81.8	**80.0**

Table 7. Performance on Daily-Omni.

	Tech & Science	Culture & Politics	Daily Life	Film & TV	Performance	Games	Sports	Music	**Overall**
OmniSeek	68.2	66.3	61.4	64.6	62.9	56.7	57.7	60.1	**62.4**

Table 8. Performance on WorldSense.

	Cartoon	Edu	Emerg	Surv	Daily	Movie	Game	Doc	Speech	Sound	Music	**Overall**
OmniSeek	61.9	74.0	56.5	67.6	61.7	48.5	60.2	47.9	53.7	61.5	61.5	**58.3**

Table 9. Performance on FutureOmni.

	Alignment	Understanding	Reasoning	$(0, 2]$min	$(2, 5]$min	**Overall**
OmniSeek	66.4	70.9	69.5	69.4	69.5	**69.5**

Table 10. Performance on OmniVideo-Test.

	SR	IMC	TCI	TA	MHR	PAR	CTI	**Overall**
OmniSeek	83.6	74.3	69.2	72.5	75.0	69.6	75.6	**74.6**

Table 11. Performance on VideoHolmes.

	STL	SPL	SOOG	SOER	SPER	MPTI	VSSR	CSA	MPO	PTG	AFA	PDP	AVDM	MESI	CRI	**Overall**
OmniSeek	77.4	59.8	74.3	87.8	47.5	79.0	87.8	44.6	78.8	61.1	71.0	72.0	71.8	84.2	87.3	**72.8**

Table 12. Performance on JointAVBench.

	Music	Sound	Speech	$(0, 1]$min	$(1, 5]$min	$(5, 10]$min	$(10, 30]$min	**Overall**
OmniSeek	41.8	51.0	47.8	56.0	49.3	45.0	42.7	**47.7**

Table 13. Performance on OmniVideoBench.

	$<$5min	5–10min	10–20min	20–30min	$>$30min	**Overall**
OmniSeek	70.2	69.6	70.1	72.5	70.5	**70.4**

Table 14. Performance on MMOU test-mini split.

	Understanding	Perception	Inference	Logical	Low	Medium	High	**Overall**
OmniSeek	46.5	43.3	45.7	38.5	54.1	42.4	35.1	**44.2**

Table 15. Performance on LVOmniBench.

	Short	Medium	Long	**Overall**
OmniSeek	84.9	78.7	71.9	**78.5**

Table 16. Performance on Video-MME.