OPEN-OMNICOACH: EVIDENCE-ENHANCED TRAJECTORY LEARNING FOR OMNI VIDEO AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Training capable omni video agents requires more than diverse question-answer pairs: agents must learn to *reliably find, connect, and correct task-relevant evidence in omni-setting*. However, we argue that existing data pipelines provide limited control over task-relevant evidence and weak supervision for these capabilities. In this work, we present **Open OmniCoach**, a new framework for evidence-seeking omni video agents. It contains a high-quality dataset, **OmniCoach-240k**, and a new agentic reinforcement learning (RL) method, Step-Level Evidence-Guided Policy Optimization (**SEPO**). For the former, OmniCoach-340k has both multi-span trajectory data and reflection trajectory data. We design a three-stage data pipeline by leveraging both fine-grained omni-evidence and the agent's own errors. Meanwhile, compared with other RL baselines, SEPO provides turn-level credit based on improvements in evidence coverage and temporal localization. Applying both trajectory data and the SEPO method, we obtain **Open Omni-Agent** models. The resulting models outperform the baselines across multiple video and audio-visual understanding tasks by significant margins, demonstrating the effectiveness of the Open OmniCoach framework. In particular, Open OmniAgent-30B improves average accuracy over Qwen3-Omni by 6.3 percentage points across nine benchmarks, with gains of 12.2 and 13.3 points on LVOmniBench and LVBench, respectively. Code, datasets, and models will release soon.

1 INTRODUCTION

Recently, agent-based approaches have been widely applied to multimodal reasoning problems (Zheng et al., 2025b; Wu et al., 2025; Hong et al., 2025b; Yang et al., 2025; Xing et al., 2026; Wang et al., 2026b;a; Meng et al., 2025), enabling active evidence acquisition and improved comprehension. In this field, high-quality task data and effective post-training methods are crucial for effective agent training (Tao et al., 2025; Xu et al., 2026; Xing et al., 2026). The former always contains interaction trajectories, while the latter has more optimization potential for better results.

Recent works extend agent-based approaches to video-audio (omni) understanding, with initial efforts constructing tool-calling interaction trajectories (Tao et al., 2025; Xu et al., 2026; Xing et al., 2026). However, these works still provide fewer details on dataset construction and offer less fine-grained evidence control for building high-quality omni-trajectories. More specifically, how can we make trajectory synthesis more controllable and fine-grained, and enable agents to learn the capabilities needed in omni settings, such as active evidence acquisition and audio-visual evidence binding? In this way, we can mitigate hallucinations in intermediate interaction steps.

To meet these requirements, we introduce **Open-OmniCoach**, a new framework for omni video agent training that includes a high-quality training dataset (**OmniCoach-240K**) and a new RL method (Step-Level Evidence-guided Policy Optimization, **SEPO**). It has *three new key features*:

1. **Generating Trajectory From Fine-Grained Omni-Evidence.** Existing omni-modal datasets (Cai et al., 2026; Geng et al., 2024) are mostly built on videos of a few minutes, rarely requiring an agent to relate distant moments or to locate brief events in long videos. Instead, we first compose evidence with task-specific skills across a hierarchy of global events and local audio-visual details, enabling question-answering and temporal grounding tasks that span diverse reasoning types and temporal scales. Then, we couple dynamic

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

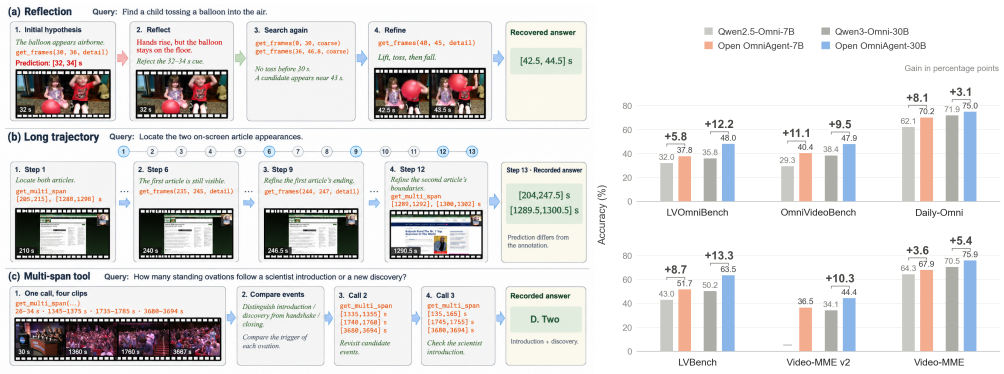

Figure 1: **Open-OmniCoach at a glance.** Left: schematic illustrations of (a) reflection that corrects an initial temporal prediction through evidence reinspection, (b) a 13-step trajectory that searches for event boundaries, and (c) multi-span tool use that jointly inspects distant clips. Right: accuracy of Open OmniAgent-7B and Open OmniAgent-30B compared with their Qwen2.5-Omni and Qwen3-Omni base models, respectively, on six audio-visual and video understanding benchmarks. Brackets indicate gains in percentage points; the dash denotes an unavailable baseline result.

evidence cropping with multi-span joint interpretation in a trajectory-generation harness, enabling frontier models to crop multiple audio-visual spans in a single tool call and jointly inspect them as reasoning unfolds.

2. **Generating Trajectory from the Agent's Own Errors.** Existing pipelines supervise agents only with successful trajectories (Xing et al., 2026; Zeng et al., 2026; Yang et al., 2025), which never show how to recover from the agent's own mistakes, although trained agents still make claims unsupported by the evidence they inspect (Qiu et al., 2026). We, instead, diagnose failed rollouts of our SFT agent at key steps against the inspected audio and frames, and let a teacher resume from the first error with an explicit correction.
3. **Crediting Rewards For Each Step.** Existing video-o3 like agents (Meng et al., 2025; Zeng et al., 2026) dierctly average all the trajectory rewards without leveraging the evidence credits for each step. Instead, we present an improved version of the RL post-training method, guided by omni-evidence.

Concretely, we present a three-stage pipeline for building our **OmniCoach-240K**. For stage 1, we build an evidence-enhanced Omni-QA pipeline. In particular, we present four steps: fine-grained evidence generation, event-and-evidence matching, skill-guided QA construction, and quality-aware filtering. Each step runs sequentially, producing refined QA data with fine-grained evidence. The first stage is a key step in generating a high-quality trajectory. For stage 2, we design a pipeline for multi-span trajectory generation, using a dynamic omni-evidence harness and omni-evidence data from stage 1 as input. We obtain 230k trajectory data in this step as multi-span trajectory data. For stage 3, we present a reflection trajectory generation process. The trained model (using stage-2 data) checks the failed rollouts with AV-decoupled stepwise verification and lets the teacher resume from the first error, yielding an extra 10k+ reflection trajectory data. We list the more detailed dataset comparison in Tab. 1.

Building on these data, we train omni-models in two stages. We first use both multi-span and reflection trajectory data for the SFT process. For the RL process, as sparse outcome rewards provide limited guidance on intermediate evidence acquisition, we introduce Step-Level Evidence-Guided Policy Optimization (SEPO), which uses the key-evidence annotations to credit each step by its incremental gain in evidence coverage and localization. Compared with the GSPO baseline, our method dynamically assigns credits to each trajectory during each step, leading to stronger performance.

Applying this recipe to open-source models (Qwen3-Omni-30B-A3B (Xu et al., 2025b) and Qwen2.5-Omni-7B (Xu et al., 2025a)), yields **Open OmniAgents, with both 30B and 7B mod-**

Table 1: Comparison of representative video and omni-video training datasets across QA diversity, dynamic evidence acquisition, and error recovery. TG: temporal grounding; MCQ: multiple-choice QA; FF: free-form QA/generation; N.R.: not reported; –: not applicable. For ours, Avg. Duration is the mean duration of unique videos covered by QA data. Multi-span denotes multiple temporal intervals requested in one tool call.

		Diverse QA				Dynamic Evidence Acquisition		Recovery Supervision
Dataset	**Mod.**	**#QA**	**Formats**	**Avg. Duration**	**Evidence Ann.**	**#Traj.**	**Multi-span**	**Error Recovery**
LongVALE (Geng et al., 2024)	AV	N.R.	TG, FF	3.9 min	Time + text	–	–	×
OmniVideo-100K (Cai et al., 2026)	AV	100K	MCQ, FF	103 s	Time + text	–	–	×
LongVT (Yang et al., 2025)	V	N.R.	TG, FF	N.R.	Time	34.6K	×	×
Video-o3 (Zeng et al., 2026)	V	N.R.	MCQ, FF	N.R.	Time	173K	×	✓
OmniAgent (Xing et al., 2026)	AV	N.R.	TG, MCQ, FF	N.R.	N.R.	58K	×	✓

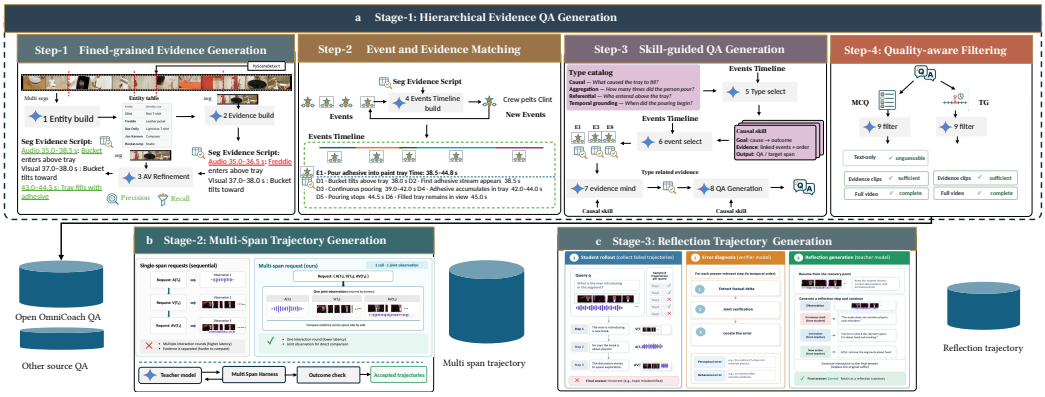

Figure 2: **OmniCoach-240K data generation pipeline. (a) Hierarchical evidence QA generation:** fine-grained audio-visual scripts are linked to a global event timeline, supporting skill-guided MCQ and TG generation followed by quality-aware filtering. **(b) Multi-span trajectory generation:** using Stage-1 and other-source QA samples, a teacher generates trajectories in the Dynamic Omni-Evidence Harness, which supports joint evidence acquisition across multiple time spans and modalities. **(c) Reflection trajectory generation:** averifier diagnoses errors in student rollouts. At confirmed recovery points, a teacher preserves the erroneous draft, adds an evidence-grounded correction and a revised action, and continues the interaction. Both teacher and reflection trajectories are retained only after passing task-level outcome checks.

els. Empirically, our Open OmniAgent-30B improves over its base model on all five audio-visual benchmarks, raising their average by 5.41 points, including +9.50 on OmniVideoBench (Li et al., 2025) and +12.23 on LVOmniBench (Tao et al., 2026). It also improves other video understanding benchmarks, e.g., +10.28 on Video-MME v2 Fu et al. (2026) and +13.26 on LVBench (Wang et al., 2024a), surpassing Qwen3-VL-235B-A22B on Video-MME v2. To verify our framework's generalization, we also train Open OmniAgent-7B from Qwen2.5-Omni-7B, which improves the audio-visual benchmarks by an average of 6.25 points. Moreover, our tasks elicit more varied evidence-seeking behaviors than OmniVideo-100K (Cai et al., 2026) at comparable video length, with 7.1 versus 5.1 effective tool-use patterns.

2 OmniCoach-240K

In this section, we present the detailed pipeline for creating our OmniCoach-240K dataset. The entire process contains three steps: Hierarchical Evidence QA Generation (Stage-1), Multi-Span Trajectory Generation (Stage-2), and Reflection Trajectory Generation (Stage-3). Each stage builds on the output of the previous stage. The final trajectory data from Stage-2 (230K+) and Stage-3 (10K+) together constitute OmniCoach-240K.

2.1 Stage-1: Hierarchical Evidence QA Generation

As shown in Fig. 2 (a), to obtain better Omni QA data, we present a new way of generating hierarchical evidence, which covers fine-grained grounded evidence across video and audio. This stage

has four steps: Hierarchical Evidence (step-1), Event and Evidence Matching (step-2), Skill-guided QA Generation (step-3), and Quality-aware Filtering (step-4).

Step-1: Fined-grained Evidence Generation. We first construct segment-level audio-visual scripts that record fine-grained evidence with time intervals and consistent entity references. Specifically, we divide each video into segments based on visual scene changes and audio pauses, and process context blocks of adjacent segments in temporal order to incrementally build a global entity table of people, sound sources, and objects. Using this table, the model generates a script for each segment, recording fine-grained facts from visual content, speech, music, and other sounds. To correct errors and recover omitted information in the initial scripts, we introduce *Audiovisual Script Refinement*. Models verify specific facts more reliably than they describe complex segments from scratch (Rawal et al., 2025; Liu et al., 2026). The model therefore uses the initial scripts and the entity table as references to re-examine the original audio-visual segments, refine event boundaries, correct sound-source attribution and audio-visual mismatches, fix factual errors, and add omitted information.

Step-2 Event and Evidence Matching. The refined scripts provide rich local details, but selecting task-relevant events and evidence directly from the full record is challenging. We therefore organize the scripts into a concise global event timeline. Using the entity table and existing event records, we integrate the scripts in temporal order, merging continuing events and adding new events and state changes. Each event retains links to its source evidence in the scripts, forming an evidence hierarchy built bottom-up that supports top-down task construction: first selecting events on the timeline, then expanding local details as needed. Appendix C.1 details script generation, refinement, and timeline construction.

Step3: Skill-guided QA Generation. Building on this evidence hierarchy, we define a skill for each MCQ reasoning type and for TG, specifying the required audio-visual cues, their relations, and how to construct the question. For example, counting requires verifying candidate events individually and aggregating those that satisfy the question's conditions, while TG targets intervals where the specified audio and visual conditions overlap in time. A skill catalog summarizes the reasoning goals and evidence requirements of each question type, serving as a concise index of the corresponding skills.

During construction, the model uses the global event timeline and skill catalog to select question types with sufficient supporting evidence and loads the corresponding skills. It then selects event combinations that satisfy the skill requirements and expands the linked fine-grained scripts to extract both answer-relevant facts and cues that distinguish easily confused entities or events. Using this evidence together with the corresponding raw audio-visual segments, the model generates QA questions and reference answers, or TG queries and target intervals. Each task retains the evidence supporting its answer or target intervals as task-level evidence annotations, used for quality filtering and reinforcement learning (Sec. 3.2). Appendix C provides the complete task skills and construction details for each task type.

Step-4: Quality-aware Filtering. For QA, we first remove samples that can be answered correctly from the question and options alone (Chen et al., 2024a). A verifier model then answers each question in two independent passes: using only the raw audio-visual segments corresponding to the annotated evidence to check whether this evidence suffices to support the reference answer, and using the full video with audio to check whether omitted events outside these segments could change the answer. We retain only samples for which both answers match the reference answer. TG follows the same verification principle, requiring a one-to-one match between predicted and target intervals with no missing or extra intervals, and a temporal IoU above a preset threshold for every matched pair. Appendices C.3 and C.4 detail the verification settings.

2.2 Stage-2: Multi-Span Trajectory Generation

As shown in Fig. 2(b), using the QA data constructed in Stage-1 together with QA samples from other sources, we generate interaction trajectories for training active audio-visual evidence acquisition. Audio-visual reasoning often requires jointly analyzing evidence distributed across time spans and modalities. Retrieving these spans through sequential requests not only increases interaction rounds but may also propagate intermediate misinterpretations or omissions into subsequent decisions, leading to error accumulation. We therefore support flexible multi-span, multimodal evidence

acquisition, allowing the model to directly compare relevant evidence while reducing unnecessary sequential interactions.

To this end, we design the *Dynamic Omni-Evidence Harness*, a multi-turn environment in which the model issues evidence requests and receives corresponding observations until it submits a final answer. Each request specifies one or more time spans and independently assigns an observation modality to each: audio (A), visual frames (V), or joint audio-visual content (AV). The harness returns the requested content together as a single observation. Compared with sequential single-span requests, multi-span requests reduce the rounds needed to retrieve the same set of spans, allowing the model to compare evidence across spans and check audio-visual correspondence within the same turn. The model can also adjust visual sampling density and spatial resolution as needed. Additionally, we adopt a cross-turn text memory mechanism similar to OmniAgent (Xing et al., 2026), retaining key information from previous observations for subsequent reasoning. Appendix D.1 details the request format and observation configuration.

We use Gemini 3.6 to generate teacher trajectories in this harness and filter them through an outcome check. For MCQ, the final selected option must be correct. For TG, both the set-level temporal intersection over union (tIoU) between predicted and target intervals and the minimum coverage across target intervals must meet their respective thresholds. We retain 230k trajectories for the initial SFT of the student model.

2.3 Stage-3: Reflection Trajectory Generation

Even after SFT, the student may still produce perceptual hallucinations or make erroneous decisions during interaction. Learning from successful trajectories alone does not necessarily equip the model to recognize and recover from its own errors. We therefore construct reflection trajectories from student failures: at confirmed error points, we introduce corrections grounded in audio-visual evidence and continue the interaction to a final answer, providing explicit supervision for error recovery.

Trajectory collection. We collect student rollouts in the same harness, sampling eight trajectories per query in the QA set and selecting failed trajectories for error diagnosis.

Error diagnosis. The verifier first selects, from the textual trajectory, key steps that affect the answer or subsequent evidence acquisition and examines them in temporal order. At each step, it extracts new or revised factual claims relative to the preceding history. Using the current observations and interaction history, it checks the audio-visual support for these claims and the student's behavioral decisions, such as whether it ignores available counter-evidence. Verification stops at the first error or inconclusive judgment; only a confirmed error defines a recovery point.

Reflection generation. At the recovery point, we preserve the preceding interaction history, current observations, and the student's erroneous draft. The teacher generates an evidence-grounded correction followed by a revised action, forming an observation $\rightarrow$ erroneous draft $\rightarrow$ correction $\rightarrow$ action step. It then continues the interaction to a final answer, replacing the original trajectory's remaining steps. We retain recoveries that pass the same task-level outcome check used in Stage-2, yielding 17K reflection trajectories. The Stage-2 teacher trajectories initialize the student through SFT, while the reflection trajectories provide additional supervision for error correction and recovery in subsequent SFT.

3 Open Omni Post-training

3.1 Supervised Fine-Tuning

We present two-round supervised fine-tuning. We first train the base model on Stage-2 multi-span trajectories with standard next-token prediction. Then, we use the Stage-3 strategy (Sec. 2.2) to turn the model's failed rollouts into reflection trajectories. We finally train the base model on both trajectory sets and use the resulting model as the start of reinforcement learning (Sec. 3.2).

3.2 Reinforcement Learning with Step-Level Credit Assignment

Base Algorithm and Reward Design. For RL, we follow Open-o3 Video (Meng et al., 2025) and use Group Sequence Policy Optimization (GSPO) (Zheng et al., 2025a) as the base optimization framework to further optimize the model's multi-turn evidence-acquisition and answering strategies. Given an audio-visual input and a question x, the reward for an interaction trajectory τ contains two items: outcome reward $r_{\text{out}}$and format reward r_{fmt},$R(x,\tau) = r_{\text{out}}(x,\tau) + r_{\text{fmt}}(\tau)$.. For the multiple-choice question (MCQ) task, the outcome reward is a binary correctness reward based on answer-option matching. For the temporal grounding (TG) task, we follow TimeLens2 (Zhu et al., 2026) to compute the outcome reward. The format reward is determined by whether the final answer follows the required output format. Then, GSPO computes a group-relative advantage A_i for each sampled trajectory. Let

$$\ell(s, A) = \min\{sA,\ \text{clip}(s, 1-\epsilon, 1+\epsilon)A\} \tag{1}$$

denotes the clipped objective term. The GSPO objective is

$$\mathcal{J}_{\text{GSPO}}(\theta) = \mathbb{E}_{x,\{\tau_i\}}\left[\frac{1}{G}\sum_{i=1}^{G}\ell(s_i(\theta), A_i)\right]. \tag{2}$$

Here, $\{\tau_i\}_{i=1}^{G}$ are G trajectories sampled by the old policy for the same input, $s_i(\theta)$ is the length-normalized sequence importance ratio, and ϵ is the clipping threshold.

Step-Level Evidence Guidance. A shared trajectory-level advantage cannot distinguish the evidence contribution of individual steps: successful trajectories may contain redundant calls, while failed trajectories may still acquire useful evidence. Ours MCQ and TG data provide evidence temporal annotations, which offer a common reference for evaluating evidence acquisition. We therefore propose Step-Level Evidence-Guided Policy Optimization (SEPO), which uses evidence-matching gains to adjust the update strength of each step.

Let $\mathcal{A}_x$ denote the annotated evidence intervals, and let $\mathcal{C}_{i,k}$ denote all candidate segments actually retrieved by trajectory τ_i up to step k. To prevent irrelevant candidates from affecting the score of the evidence already acquired, the scorer performs a bounded exhaustive subset search over $\mathcal{C}_{i,k}$ and denotes the highest-scoring subset by $\mathcal{S}^{\star}_{i,k}$:

$$\mathcal{S}^{\star}_{i,k} \in \arg\max_{\mathcal{S}\subseteq\mathcal{C}_{i,k}} q(\mathcal{S}, \mathcal{A}_x). \tag{3}$$

This search is used only to compute the evidence utility and does not represent an additional model action. The current evidence utility is

$$\widehat{U}_{i,k} = q(\mathcal{S}^{\star}_{i,k}, \mathcal{A}_x). \tag{4}$$

We maintain the historical best utility and use its improvement as the evidence-matching gain of the current step:

$$U_{i,k} = \max\left(U_{i,k-1}, \widehat{U}_{i,k}\right), \qquad \Delta U_{i,k} = U_{i,k} - U_{i,k-1}, \qquad U_{i,0} = 0. \tag{5}$$

We standardize and clip the gains within each trajectory, and define the step-level advantage as

$$\widetilde{A}_{i,k} = A_i + \lambda|A_i|\,\text{clip}\left(\frac{\Delta U_{i,k} - \mu_i}{\sigma_i + \delta}, -1, 1\right). \tag{6}$$

Here, μ_i and σ_i are the mean and standard deviation of the step gains within trajectory i, $\delta > 0$ is a numerical stability term, and $0 \le \lambda < 1$ controls the adjustment strength. This adjustment preserves the sign of the advantage, giving high-gain steps stronger reinforcement in positive-advantage trajectories and weaker suppression in negative-advantage trajectories.

We adopt the GSPO-token formulation, assign each step advantage to its corresponding generated tokens, and maximize the following objective (Zheng et al., 2025a):

$$\mathcal{J}_{\text{SEPO}}(\theta) = \mathbb{E}_{x,\{\tau_i\}}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{N_i}\sum_{t=1}^{N_i}\ell\Big(s_{i,t}(\theta), \widetilde{A}_{i,k(i,t)}\Big)\right]. \tag{7}$$

Here, N_i is the number of generated tokens in trajectory i, and $k(i,t)$ denotes the step to which token t belongs. The value of $s_{i,t}(\theta)$ is equal to $s_i(\theta)$, while gradients propagate only through the conditional generation probability of the current token. Thus, the step advantage adjusts the gradient contribution of the corresponding tokens. When $\lambda = 0$, the objective reduces to the GSPO update.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Evaluation Benchmarks. We evaluate our method on benchmarks covering omni-video (audio-visual) understanding and general video understanding. WorldSense (Hong et al., 2025a), Daily-Omni (Zhou et al., 2025), JointAVBench (Chao et al., 2025), OmniVideoBench (Li et al., 2025), and LVOmniBench (Tao et al., 2026) assess audio-visual understanding, focusing on the integration of auditory and visual evidence. Video-MME (Fu et al., 2024), Video-MME v2 (Fu et al., 2026), LongVideoBench (Wu et al., 2024), LVBench (Wang et al., 2024a), and VideoZeroBench (Meng et al., 2026) assess general video understanding, including temporal reasoning.

Baselines and Comparison Models. We compare against three groups of baselines. Proprietary models include GPT-4o (OpenAI, 2024), Gemini-2.0-Flash[1], Gemini-2.5-Pro (Gemini Team, 2025), and Gemini-3.0-Pro (Tao et al., 2026). Open-source models contains VideoLLaMA2 (Cheng et al., 2024), Qwen2.5-VL (Bai et al., 2025), Qwen2.5-Omni (Xu et al., 2025a), and Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b). For agentic models, we compare with OmniAgent (Xing et al., 2026) and Video-o3 (Zeng et al., 2026).

Training details and inference. We build our agent on Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b). As discussed in Sec. 3, we first use 230K multi-span trajectories. Then, we combine 230K seed trajectories with 17K additional rollouts, yielding 247K training examples. Both supervised training runs use a learning rate of 10^{-5} for two epochs. We use AdamW as the optimizer. For RL training, we adopt our SEPO, with 8 rollouts per task and a learning rate of 10^{-6} for 834 updates. At inference, we allow up to 12 interaction turns. We also train Open OmniAgent-7B from Qwen2.5-Omni 7B (Xu et al., 2025a) with the same recipe.

4.2 MAIN RESULTS

Results on omni-video understanding benchmarks. As shown in Tab. 2, our Open OmniAgent-30B achieves the best reported performance among the compared open-source models on all five audio-visual benchmarks, with an average accuracy of 58.1%, exceeding the reported Qwen3-Omni result by 5.4 percentage points. The largest gains are observed on OmniVideoBench and LVOmniBench, reaching 9.5 and 12.2 percentage points, respectively. These gains are consistent with our interaction-oriented supervision, which trains agents to acquire and integrate task-relevant evidence through multi-turn interactions. The improvements also extend to Qwen2.5-Omni-7B: Open OmniAgent-7B achieves an average accuracy of 51.3%, improving over the reported base-model result by 6.3 percentage points. These consistent gains across model backbones support the transferability of our training data and recipe.

Results on general video understanding benchmarks. Open OmniAgent-30B also achieves competitive results on general video understanding benchmarks (Tab. 3), improving over the reported Qwen3-Omni results by 10.3 and 13.3 percentage points in accuracy on Video-MME v2 and LVBench, respectively. These results suggest that learning to acquire and integrate task-relevant evidence through interaction also benefits broader video understanding, beyond dedicated audio-visual benchmarks.

4.3 ABLATION STUDIES AND ANALYSIS

Scaling Effect of Open-OmniCoach. The OmniVideo-100K (Cai et al., 2026) provides question–answer pairs, whereas our data additionally provides interaction trajectories. We therefore train on each dataset in the supervision format it provides: direct SFT, which trains the model to answer from the input media without tool calls, for OmniVideo-100K, and agentic SFT, which trains the model on multi-turn interaction trajectories, for our data. We vary the training data from 25% to 100% of each dataset and report the mean accuracy on WorldSense, Daily-Omni, OmniVideoBench, and LVOmniBench (Fig. 3). We list more detailed results in Tab. 7. Overall, our results show that our data pipeline scales better and performs more effectively than OmniVideo-100k.

[1] https://deepmind.google/models/model-cards/

Table 2: Results on Omni-Video understanding benchmarks. All scores are accuracy (%). Bold and underline mark the best and second-best open-source results; – denotes unavailable results. Avg. is the mean over the five benchmarks, and Δ is computed against the reported scores of each base model: Qwen2.5-Omni for Open OmniAgent-7B and Qwen3-Omni for Open OmniAgent-30B.

Method	Size	WorldSense	Daily-Omni	JointAVBench	OmniVideoBench	LVOmniBench	Avg.
Proprietary Models							
GPT-4o	–	42.6	56.5	45.0	–	–	–
Gemini-2.0-Flash	–	–	67.8	–	41.5	42.9	–
Gemini-2.5-Pro	–	65.1	80.6	65.3	58.9	–	–
Gemini-3.0-Pro	–	66.4	81.1	–	61.8	65.8	–
Open-source Generalist Models							
video-SALMONN 2+	7B	50.9	71.8	46.7	36.4	32.7	47.7
OmniVinci	9B	48.2	66.5	–	32.1	–	–
Qwen2.5-VL	7B	35.5	40.7	47.7	29.8	–	–
Qwen2.5-Omni	7B	45.4	62.1	56.5	29.3	32.0	45.1
Qwen3-Omni-Instruct	30B-A3B	<u>54.0</u>	<u>71.9</u>	<u>63.6</u>	38.4	35.8	<u>52.7</u>
Open-source Agentic Models							
OmniAgent	7B	47.2	64.8	–	37.1	–	–
Open OmniAgent (Ours)	7B	48.9	70.2	59.3	<u>40.4</u>	<u>37.8</u>	51.3
Δ *vs. Qwen2.5-Omni*		+3.5	+8.1	+2.8	+11.1	+5.8	+6.3
Open OmniAgent (Ours)	30B-A3B	**54.8**	**75.0**	**64.9**	**47.9**	**48.0**	**58.1**
Δ *vs. Qwen3-Omni*		+0.8	+3.2	+1.3	+9.5	+12.2	+5.4

Table 3: Results on vision-centric video understanding benchmarks. Scores are accuracy (%) except for the official Video-MME v2 Rating; for VideoZeroBench (VZB), we report Level-3 accuracy. Video-MME v2 baselines use the with-subtitle/audio setting of its leaderboard. Other conventions follow Tab. 2.

Method	Size	Video-MME	Video-MME v2		LongVideoBench	LVBench	VZB
			Acc	Rating			L3
Proprietary Models							
GPT-4o	–	71.9	–	–	66.7	48.9	–
Gemini-1.5-Pro	–	75.0	–	–	64.0	33.1	–
Gemini-2.5-Pro	–	–	–	–	76.8	67.4	16.4
Gemini-3.0-Pro	–	88.6	66.1	49.4	75.9	77.0	17.0
Open-source Generalist Models							
Qwen2.5-Omni	7B	64.3	–	–	53.8	43.0	–
Qwen3-Omni-Instruct	30B-A3B	70.5	34.1	17.1	<u>66.9</u>	50.2	–
Qwen3-VL-Instruct	8B	71.4	35.7	18.2	66.4	58.0	8.2
Qwen3-VL-Instruct	235B-A22B	<u>79.2</u>	<u>43.3</u>	<u>25.0</u>	–	<u>67.7</u>	9.6
Qwen3.5	397B-A17B	**83.7**	42.1	24.5	–	**75.5**	<u>10.6</u>
Open-source Agentic Models							
OmniAgent	7B	67.8	–	–	–	50.5	–
Open OmniAgent-7B	7B	67.9	36.5	18.4	57.1	51.7	6.4
Δ *vs. Qwen2.5-Omni*		+3.6			+3.3	+8.7	
Open OmniAgent-30B	30B-A3B	75.9	**44.4**	**25.5**	**67.4**	63.5	**12.0**
Δ *vs. Qwen3-Omni*		+5.4	+10.3	+8.4	+0.5	+13.3	

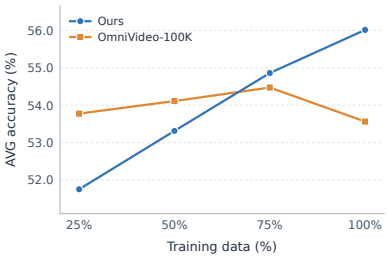

Figure 3: Data scaling. Performance using 25%–100% of OmniVideo-100K and our dataset. Per-benchmark results are in Appendix Tab. 7.

Table 4: Query-source comparison with 5K trajectories per source, generated using the same procedure. Accuracy (%).

Query source	Daily-Omni	LVOmni Bench	OmniVideo Bench	World Sense
OmniVideo-100K	67.4	39.4	41.5	50.4
Ours	**68.4**	**40.3**	**42.2**	**51.8**

Table 5: Ablation studies on Qwen3-Omni-30B-A3B-Instruct. Each row adds one component to the row above. The gray row marks the full recipe. Accuracy (%) ↑.

Setting	**Video understanding**		**Audio-visual reasoning**	
	Video-MME	Video-MME-v2	OmniVideoBench	LVOmniBench
Qwen3-Omni-30B-A3B-Instruct	70.5	34.1	38.4	35.8
+ Muti-Span Trajectory	72.9	40.7	44.7	46.8
+ Reflection Trajectory	74.4	42.6	46.3	47.5
+ GSPO (RL baseline)	75.4	43.3	46.9	47.7
+ SEPO (Our Method, final)	75.9	44.4	47.9	48.0

Seed and Reflection are trained independently from the base checkpoint; Reflection adds reflection trajectories to the seed trajectories. GSPO and SEPO both start from the Reflection checkpoint.

Effectiveness of Query Set in Open-OmniCoach. Our query set includes long-video tasks and offers broader task diversity (Tab. 1). To evaluate its value for agent training, we compare training on 5K trajectories generated from each query source using the same procedure. Using our queries instead of OmniVideo-100K queries improves accuracy across all four benchmarks by an average of 1.0 percentage point (Tab. 4).

Effectiveness of Open-OmniCoach Framework. We list the detailed effectiveness of Open-OmniCoach in Tab. 5. Starting from the baseline model, adding multi-space trajectories leads to a significant improvement, as shown in the second row. With reflection trajectories, we observe more improvements across all four benchmarks in the third row. For the strong GSPO baseline, adopting RL post-training, we still find further average 1.0% improvements across four benchmarks. Then, using our proposed SEPO, the further gains add 1.10 and 0.99 points over GSPO on Video-MME-v2 and OmniVideoBench.

5 Conclusion

In this work, we introduced Open-OmniCoach, a framework for training evidence-seeking omni video agents through evidence-enhanced trajectory learning. We construct fine-grained audio-visual evidence annotations and use them to generate multi-span interaction trajectories that teach agents to actively acquire and connect evidence across time and modalities. We further turn the agent's own failed rollouts into reflection trajectories, providing explicit supervision for recovering from intermediate errors. Building on these data, we propose Step-Level Evidence-Guided Policy Optimization (SEPO), which uses incremental evidence-matching gains to provide finer-grained credit for individual interaction steps. Across both 7B and 30B backbones, Open OmniAgent consistently improves audio-visual reasoning and general video understanding, while our ablations demonstrate the complementary benefits of multi-span trajectories, reflection supervision, and step-level evidence guidance. These results highlight the importance of explicitly supervising how omni video agents find, connect, and correct task-relevant evidence.

AI USE STATEMENT

Generative AI models are used as an integral part of the research pipeline described in this paper, including audio-visual annotation and refinement. We also used generative AI tools during the research and manuscript preparation process for brainstorming, language editing, and figure prototyping. All generated training data are subject to the verification and quality-filtering procedures described in the paper. All technical claims, experimental results, analyses, references, and final manuscript content were reviewed and verified by the authors. The authors take full responsibility for the content of this work.

REPRODUCIBILITY STATEMENT

The data construction section describes audio-visual annotation and refinement. The Agentic post-training section specifies the supervised training procedure. Training configurations and evaluation protocols are further documented in the appendices. We plan to release the dataset, training and evaluation code, and trained model checkpoints to facilitate reproduction and future research.

REFERENCES

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, 2025. URL https://arxiv.org/abs/2502.13923.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pp. 1708–1718, 2021.

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, 2026. URL https://arxiv.org/abs/2606.14702v2.

Jianghan Chao, Jianzhang Gao, Wenhui Tan, Yuchong Sun, Ruihua Song, and Liyun Ru. JointAVBench: A Benchmark for Joint Audio-Visual Reasoning Evaluation, 2025. URL https://arxiv.org/abs/2512.12772.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024a. URL https://arxiv.org/abs/2403.20330.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-Wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and S. Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *CVPR*, pp. 13320–13331, 2024b.

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, 2024. URL https://arxiv.org/abs/2406.07476.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis, 2024. URL https://arxiv.org/abs/2405.21075.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, Yongkang Xie, Xiawu Zheng, Xue Yang, Haoyu Cao, Yunsheng Wu, Ziwei Liu, Xing Sun, Caifeng Shan, and Ran He. Video-MME-v2: Towards the Next Stage in Benchmarks for Comprehensive Video Understanding, 2026. URL https://arxiv.org/abs/2604.05015.

Gemini Team. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. Technical report, Google, 2025. URL https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf.

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos, 2024. URL https://arxiv.org/abs/2411.19772v2.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. WorldSense: Evaluating Real-world Omnimodal Understanding for Multimodal LLMs, 2025a. URL https://arxiv.org/abs/2502.04326.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. Deepeyesv2: Toward agentic multimodal model, 2025b. URL https://arxiv.org/abs/2511.05271v1.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Wentao Wang, Zhenghao Song, Dingling Zhang, Ying He, Haoxiang Liu, Yuxuan Wang, Qiufeng Wang, Jiafu Tang, Zhenhe Wu, Jiehui Luo, Zhiyu Pan, Weihao Xie, Chenchen Zhang, Zhaohui Wang, Jiayi Tian, Yanghai Wang, Zhe Cao, Minxin Dai, Ke Wang, Runzhe Wen, Yinghao Ma, Yaning Pan, Sungkyun Chang, Termeh Taheri, Haiwen Xia, Christos Plachouras, Emmanouil Benetos, Yizhi Li, Ge Zhang, Jian Yang, Tianhao Peng, Zili Wang, Minghao Liu, Junran Peng, Zhaoxiang Zhang, and Jiaheng Liu. OmniVideoBench: Towards Audio-Visual Understanding Evaluation for Omni MLLMs, 2025. URL https://arxiv.org/abs/2510.10689.

Wenjie Li, Yujie Zhang, Haoran Sun, Xingqi He, Hongcheng Gao, Chenglong Ma, Ming Hu, Guankun Wang, Shiyi Yao, Renhao Yang, Hongliang Ren, Lei Wang, Junjun He, and Yankai Jiang. Medscope: Incentivizing "think with videos" for clinical reasoning via coarse-to-fine tool calling, 2026. URL https://arxiv.org/abs/2602.13332v1.

Jingyang Lin, Jialian Wu, Jiang Liu, Ximeng Sun, Ze Wang, Xiaodong Yu, Jiebo Luo, Zicheng Liu, and Emad Barsoum. Videoseek: Long-horizon video agent with tool-guided seeking, 2026. URL https://arxiv.org/abs/2603.20185.

Weisong Liu, Haochen Wang, Kuan Gao, Yuhao Wang, Yikang Zhou, Zhongwei Ren, Jacky Mai, Anna Wang, Yanwei Li, Jason Li, and Zhaoxiang Zhang. MotionAtlas: Detailed region captioning for motion-centric videos, 2026. URL https://arxiv.org/abs/2606.29531.

Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, Eng Siong Chng, and Xie Chen. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception, 2025. URL https://arxiv.org/abs/2510.12720v1.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman H. Khan, and Fahad Shahbaz Khan. Videochatgpt: Towards detailed video understanding via large vision and language models. In *ACL*, 2023.

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. *arXiv preprint arXiv:2510.20579*, 2025.

Jiahao Meng, Tan Yue, Qi Xu, Haochen Wang, Zhongwei Ren, Weisong Liu, Yuhao Wang, Renrui Zhang, Yunhai Tong, and Haodong Duan. VideoZeroBench: Probing the Limits of Video MLLMs with Spatio-Temporal Evidence Verification, 2026. URL https://arxiv.org/abs/2604.01569.

OpenAI. GPT-4o System Card. Technical report, OpenAI, 2024. URL https://openai.com/index/gpt-4o-system-card/.

Chenhao Qiu, Yechao Zhang, Xin Luo, Shien Song, and Xusheng Liu. VideoSEAL: Mitigating evidence misalignment in agentic long video understanding by decoupling answer authority, 2026. URL https://arxiv.org/abs/2605.12571.

Ruchit Rawal, Reza Shirkavand, Heng Huang, Gowthami Somepalli, and Tom Goldstein. ARGUS: Hallucination and omission evaluation in video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Rawal_ARGUS_Hallucination_and_Omission_Evaluation_in_Video-LLMs_ICCV_2025_paper.pdf.

Keda Tao, Wenjie Du, Bohan Yu, Weiqiang Wang, Jian Liu, and Huan Wang. Active perception agent for omnimodal audio-video understanding, 2025. URL https://arxiv.org/abs/2512.23646v2.

Keda Tao, Yuhua Zheng, Jia Xu, Wenjie Du, Kele Shao, Hesong Wang, Xueyi Chen, Xin Jin, Junhan Zhu, Bohan Yu, Weiqiang Wang, Jian Liu, Can Qin, Yulun Zhang, Ming-Hsuan Yang, and Huan Wang. LVOmniBench: Pioneering Long Audio-Video Understanding Evaluation for Omnimodal LLMs, 2026. URL https://arxiv.org/abs/2603.19217.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Sule Bai, Zijian Kang, Jiashi Feng, Zhuochen Wang, et al. Traceable evidence enhanced visual grounded reasoning: Evaluation and method. In *International Conference on Learning Representations*, volume 2026, pp. 148769–148794, 2026a.

Jiacong Wang, Zijian Kang, Haochen Wang, Xiao Liang, Ya Wang, Jiawen Li, Bohong Wu, Jiao Ran, Haiyong Jiang, Chao Feng, et al. Vgr: Visual grounded reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 89506–89528, 2026b.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. LVBench: An Extreme Long Video Understanding Benchmark, 2024a. URL https://arxiv.org/abs/2406.08035.

Yanghai Wang, Jiahao Wang, Jiafu Tang, Yuanxing Zhang, Zhe Cao, Hanyan Bian, Zijie Zhang, Weiliang Luo, Zhiyu Pan, Zixuan Dong, Jiaheng Liu, and Zhaoxiang Zhang. Avscap: Orchestrating audio-visual synergy for omni-modal video captioning, 2026c. URL https://arxiv.org/abs/2607.12820v1.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *ICLR*, 2024b.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding, 2024. URL https://arxiv.org/abs/2407.15754.

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. MMSearch-R1: Incentivizing LMMs to search, 2025. URL https://arxiv.org/abs/2506.20670.

Zhenghao Xing, Ruiyang Xu, Yuxuan Wang, Jinzheng He, Ziyang Ma, Qize Yang, Yunfei Chu, Jin Xu, Junyang Lin, Chi-Wing Fu, and Pheng-Ann Heng. Native active perception as reasoning for omni-modal understanding, 2026. URL https://arxiv.org/abs/2606.19341.

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

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025b. URL https://arxiv.org/abs/2509.17765.

Ke Xu, Yuhao Wang, Ziyang Cheng, Hongcheng Liu, Yanfeng Wang, and Yu Wang. Agentic active omni-modal perception for multi-hop audio-visual reasoning, 2026. URL https://arxiv.org/abs/2605.28192.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, and Lidong Bing. Longvt: Incentivizing "thinking with long videos" via native tool calling, 2025. URL https://arxiv.org/abs/2511.20785.

Linli Yao, Yuancheng Wei, Yaojie Zhang, Lei Li, Xinlong Chen, Feifan Song, Ziyue Wang, Kun Ouyang, Yuanxin Liu, Lingpeng Kong, Qi Liu, Pengfei Wan, Kun Gai, Yuanxing Zhang, and Xu Sun. Timechat-captioner: Scripting multi-scene videos with time-aware and structural audio-visual captions, 2026. URL https://arxiv.org/abs/2602.08711v1.

Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, Ziang Yan, Yi Wang, Hongjie Zhang, Yali Wang, and Limin Wang. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning, 2026. URL https://arxiv.org/abs/2601.23224v2.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *TMLR*, 2024.

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, 2025a. URL https://arxiv.org/abs/2507.18071.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning, 2025b. URL https://arxiv.org/abs/2505.14362v1.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-Omni: Towards Audio-Visual Reasoning with Temporal Alignment across Modalities, 2025. URL https://arxiv.org/abs/2505.17862.

Yuhan Zhu, Changlian Ma, Xiangyu Zeng, Xinhao Li, Zhiqiu Zhang, Songze Li, Jun Zhang, Tianxiang Jiang, Yuandong Yang, Ziang Yan, Zikang Wang, Xinyu Chen, Haoran Chen, Shaowei Zhang, and Limin Wang. TimeLens2: Generalist video temporal grounding with multimodal llms, 2026. URL https://arxiv.org/abs/2607.17423.

Overview. This appendix describes the evidence hierarchy used for task construction, the interaction protocol used for trajectory synthesis, the verification and recovery procedure, and the evidence utility used for step-level policy optimization. It also reports additional experiments, including a temporal-grounding comparison and controlled experiments moved out of the main text. In addition, the appendix C.4 gives the grounding results, while Appendices C–E provide implementation details. We also discuss the future work of our data pipeline and methods.

A Related Work

A.1 From Vision-Language to Omni-Modal Datasets

Large-scale video-language datasets have laid the foundation for video multimodal models. WebVid (Bain et al., 2021), InternVid (Wang et al., 2024b), and Panda-70M (Chen et al., 2024b) scale video-text pretraining with large-scale web videos and automatically generated captions. Subsequent efforts such as Video-ChatGPT (Maaz et al., 2023) and LLaVA-Video (Zhang et al., 2024) further introduce instruction-following data for video understanding. However, these datasets remain predominantly vision-centric and provide limited supervision for rich audio-visual interactions.

Recent efforts extend video supervision toward omni-modal understanding by explicitly incorporating audio and its interaction with visual content. Such datasets, including Omni-Captioner (Ma et al., 2025), TimeChat-Captioner (Yao et al., 2026), and AVSCap (Wang et al., 2026c), increasingly organize audio-visual information into structured representations and construct supervision that captures cross-modal and temporal interactions. OmniVideo-100K (Cai et al., 2026) builds entity-anchored structured audio-visual scripts and mines cross-segment multimodal evidence chains to synthesize reasoning-oriented question-answer pairs, improving the consistency and grounding of long-video supervision. However, existing omni datasets mainly use evidence to construct task-level supervision, leaving the agent's evidence-seeking process unsupervised. Our trajectory dataset further supervises dynamic evidence acquisition and error recovery, enabling agents to learn how to find, connect, and correct task-relevant evidence.

A.2 From Video MLLMs to Omni Agents

Vision-language models (VLMs), such as Qwen2.5-VL (Bai et al., 2025), have established strong multimodal perception and reasoning capabilities across images and videos. Building on these advances, Video MLLMs have evolved from visual-centric understanding toward unified audio-visual perception. VideoLLaMA2 (Cheng et al., 2024) incorporates spatial-temporal modeling and audio understanding, while recent omni-modal models, including Qwen2.5-Omni (Xu et al., 2025a) and Qwen3-Omni (Xu et al., 2025b), further integrate visual, acoustic, and linguistic information for joint audio-visual reasoning.

Beyond passive perception, recent works increasingly formulate video understanding as an active evidence-seeking process. LongVT (Yang et al., 2025), Open-o3-Video (Meng et al., 2025), VideoSeek (Lin et al., 2026), and MedScope (Li et al., 2026) enable models to iteratively retrieve or inspect task-relevant video evidence through tool use. This paradigm has further extended to omni-modal agents: OmniAgent (Tao et al., 2025; Xing et al., 2026) introduces active perception for audio-visual understanding, while AOP-Agent (Xu et al., 2026) targets multi-hop audio-visual reasoning through agentic perception. However, these works mainly focus on agent architectures and interaction mechanisms, with limited supervision for how evidence should be acquired, integrated, and corrected throughout the reasoning process. Our work addresses this gap with hierarchical supervision over evidence-grounded tasks, interaction trajectories, and error recovery. By building trajectories on stronger Omni-QA data, we obtain high-quality omni-training data.

A.3 Reinforcement Learning for Multimodal Agents

Reinforcement learning has increasingly been used to elicit active perception and tool-use behaviors in multimodal models. DeepEyes (Zheng et al., 2025b) learns to invoke visual tools through end-to-end reinforcement learning, while MMSearch-R1 (Wu et al., 2025) optimizes multi-turn multimodal search with outcome-based rewards. Related work on visual grounded reasoning further incorporates explicit evidence signals into reinforcement learning to improve the faithfulness of intermediate

Table 6: Temporal grounding. TimeLens and LongVALE report mIoU (%). Baseline TimeLens scores are sample-weighted over the three TimeLens-Bench subsets reported by TimeLens2 (Qwen2.5-VL-7B by TimeLens) (Zhu et al., 2026); LongVALE baselines follow the LongVALE report and OmniAgent (Geng et al., 2024; Xing et al., 2026); – denotes unreported results.

Method	TimeLens mIoU (%) ↑	LongVALE mIoU (%) ↑
Proprietary Models		
Gemini-2.5-Pro	58.2	–
Open-source Generalist Models		
Qwen2.5-VL-7B	34.3	–
Qwen3-VL-8B	49.6	–
Qwen3-VL-235B-A22B	<u>52.7</u>	–
Qwen3.5-397B-A17B	**53.5**	–
Qwen3-Omni-30B-A3B	–	26.1
Open-source Agentic Models		
OmniAgent-7B	–	39.1
Open OmniAgent-7B	45.9	<u>43.2</u>
Open OmniAgent-30B	49.8	**47.5**

Table 7: Per-benchmark data scaling. Models trained on OmniVideo-100K use direct SFT, while models trained on Open OmniCoach use agentic SFT. Avg. is the unweighted mean over the four benchmarks. Accuracy (%) ↑.

Training data	WorldSense	Daily-Omni	OmniVideoBench	LVOmniBench	Avg.
OmniVideo-100K, direct SFT					
25%	54.98	73.02	45.10	42.00	53.78
50%	54.82	73.02	45.80	42.80	54.11
75%	54.70	75.10	44.80	43.29	54.47
100%	52.84	72.51	46.40	42.50	53.56
Open OmniCoach, agentic SFT					
25%	52.60	69.30	42.70	42.41	51.75
50%	53.33	71.21	44.43	44.28	53.31
75%	54.06	73.11	46.17	46.10	54.86
100%	54.79	75.02	47.03	47.21	56.01

reasoning (Wang et al., 2026a). At the optimization level, GSPO (Zheng et al., 2025a) provides stable sequence-level policy optimization for large models. However, these approaches largely rely on trajectory-level outcomes or static grounding rewards and provide limited credit assignment for individual evidence-acquisition steps. Our SEPO instead derives step-level evidence gains from task annotations and uses them to reweight trajectory-level advantages, providing finer-grained supervision for multi-turn evidence seeking.

B Additional Experiments and Analysis

Temporal grounding. Table 6 reports mIoU on TimeLens and LongVALE. Open OmniAgent-7B obtains 45.87% on TimeLens and 43.23% on LongVALE, while Open OmniAgent-30B obtains 49.75% and 47.52%, respectively. The table keeps the two benchmarks separate because their annotations and evaluation protocols are different; unavailable results are marked as "–".

Per-benchmark data scaling. Table 7 gives the per-benchmark values behind Fig. 3; cells without a verified run remain explicit placeholders until the corresponding evaluation is completed.

C Task Synthesis Details

C.1 Evidence Construction

An initial pass over coarse video chunks establishes a global entity table. Audio and visual candidate boundaries are combined to divide the video into segments. A VLM annotates speech, music, non-speech sounds, and visual content, allowing their temporal intervals to overlap. Each observation records its interval, evidence type, description, and entity references. The examiner revisits the original media with the draft annotations and produces a revised annotation. Entity references are reconciled across segments, and unresolved identity assignments are removed.

Let $\mathcal{E}^{(t)}$ denote the corrected observations from segment t. Starting from an empty timeline, we update the event representation sequentially:

$$\mathcal{H}_t = \mathrm{Update}\big(\mathcal{H}_{t-1}, \mathcal{E}^{(t)}, \mathcal{G}\big). \tag{8}$$

The update preserves earlier entries and may revise the latest entry when the next segment continues the same event. New events are appended with references to their supporting observations. Observations that do not create a new timeline entry remain available for later retrieval beyond the compressed event overview.

C.2 Task Planning and Evidence Expansion

Task skills specify the required operation, candidate-selection rules, and supporting evidence. For QA, the catalog contains 18 reasoning skills. Candidate generation proceeds in two rounds, both using the complete timeline. The second round additionally receives the first candidate bank and revises it to address omissions, incomplete candidates, and task mismatch. For temporal grounding, candidate planning identifies audiovisual conditions suitable for localization, including conditions that recur across events.

Each candidate is associated with relevant events and potentially confounding events. Expansion retrieves their explicitly referenced observations and eligible observations overlapping their temporal windows. Evidence identifiers link the retrieved details to the underlying annotations. Timeline summaries guide retrieval, while the detailed observations support task formulation. Grounding additionally requires a full-video search after the query is fixed, because the planning timeline may compress repeated occurrences.

Task planning produces a task specification and references to candidate events. The final instruction and answer are constructed after the relevant local evidence has been expanded. Task-specific requirements guide both the selection of evidence and the form of the resulting task.

Each generated task is accompanied by its supporting evidence. Before accepting a sample, we ask a model to answer the final instruction from this evidence and compare its answer with the reference answer under task-specific criteria. The retained evidence intervals and modalities define $\mathcal{A}_x$.

C.3 Question Answering

The question generator uses the selected evidence, relevant entities, and task requirements to formulate a draft question. A separate solver answers the draft question from the corresponding original audio and video without annotation captions, a reference answer, or answer options. Five distractors are then constructed to form a six-option question.

The draft undergoes a text-only audit for ambiguity, answer leakage, and inconsistent option construction, followed by a media-based revision using the selected and confounding events. This revision may update the question, options, answer, and task label. A fresh solver then answers the revised question without seeing the options. A separate text-only assessment compares this answer with every option without access to the correct label. The sample is accepted only when exactly one option matches and it agrees with the label.

The full-video check additionally verifies that omitted events do not change the reference answer. Final samples retain source-segment references and their association with the original draft.

C.4 Temporal Grounding

For temporal grounding, planning identifies a candidate event or a recurring audiovisual relation. We expand the relevant observations to establish the entities, audible and visible conditions, and potentially confounding events. These details guide the formulation of a grounding query and the construction of its temporal answer. Once the query has been specified, we search the complete video evidence for all matching occurrences. Each occurrence is linked to supporting observations, from which its reference interval is derived.

For an occurrence supported by intervals $\{[s_i, e_i]\}$, its temporal label is constructed as their common intersection:

$$[s^*, e^*] = \left[\max_i s_i,\ \min_i e_i\right]. \tag{9}$$

Each occurrence requires both audio and visual evidence with a non-empty common intersection. An independent ledger audit checks semantic consistency, evidence support, and completeness of the occurrence set.

Media verification then scans the full video in bounded batches. A solver receives the query and media without reference intervals and predicts all matching occurrences. Acceptance requires successful localization with both modalities and unsuccessful localization with either modality alone. Localization is evaluated by matching predicted and reference occurrences one-to-one, requiring no unmatched occurrences and a temporal IoU above the specified threshold for every matched pair.

C.5 Additional Task Formats

These formats are auxiliary outputs used during construction and are not counted as final QA or temporal-grounding tasks. For captioning, consecutive segments are grouped into coherent stories, and their detailed observations are used to generate descriptions. Cloze tasks mask selected spans in these descriptions; blanks recoverable by a text-only solver are removed. Timeline-description tasks organize the video into chronological event descriptions, with checks for temporal coverage, non-overlap, and support from the evidence annotations. These output timelines are distinct from the internal timeline used for task planning.

D Trajectory Synthesis Details

D.1 Tool Interface and Interaction Protocol

In the main-text notation, $\mathrm{MultiSpan}$ denotes the `get_multi_span` tool. The ordered list specifies each requested interval and modality; sampling and resolution arguments are omitted from the main-text equations. The agent generates the textual observation o_t, whereas u_t contains the media supplied by the environment.

Each trajectory starts with the task instruction, tool specifications, timestamped frames covering the video, and the full audio track when available. Initial frames are sampled at 0.5 FPS. If this produces more than 256 candidates, we uniformly subsample them to 256 frames. Each initial frame uses a pixel budget of 224^2.

At each step, the agent produces a JSON object containing `observation`, `think`, and `action`. The observation describes the latest available media, while the reasoning combines relevant evidence and determines the next action. The agent either invokes one tool or terminates with an `answer` action. Direct answers are allowed at the first step, and the default interaction budget is 32 decision steps. After tool use, observations include temporal references in `[MM:SS.m,MM:SS.m]` format. Table 8 summarizes the available tools and their per-call limits.

For visual retrieval, `level=coarse` and `level=detail` sample frames at 1 and 2 FPS, respectively. Frames preserve their aspect ratio, with the longest side limited to 384 pixels by default or 1280 pixels under `resolution=high`. High-resolution intervals are limited to 5 seconds, and a multi-span call may contain at most two such intervals. Each multi-span entry also follows the limits of its corresponding single-interval tool. Tool boundaries use `MM:SS.m` with one decimal place and must lie within the video; requested audio intervals must last at least 0.1 seconds. For videos without audio, `get_audio` is unavailable and `get_clip` returns frames alone.

Table 8: Tool outputs and per-call limits.

Tool	Intervals	Returned evidence	Per-call limits
`get_frames`	1	Sampled frames	At most 256 frames
`get_audio`	1	Continuous audio	At most 256 seconds
`get_clip`	1	Frames and audio from the same interval	At most 60 seconds
`get_multi_span`	2–8 (ordered)	Frames, audio, or both per interval	At most 256 frames and 128 seconds of audio in total

Multi-span results preserve the requested order and identify each interval as `span_1`, `span_2`, and so on. The agent is instructed to describe the evidence or uncertainty for every returned interval and then compare them. After a successful retrieval, subsequent model inputs retain raw media from the latest result while preserving earlier textual history. All intervals returned by a multi-span call remain visible together. Earlier media can be inspected again through a new tool call.

D.2 Perceptual and Agentic Verification

Perceptual verification audits newly introduced or revised claims, as well as claims repeated despite contradictory evidence. Earlier observations provide context for identifying these changes; factual support comes from the media available at the current step. Audio claims are checked against audio, visual claims against frames, and cross-modal associations against joint audio-visual evidence. We distinguish unsupported claims from claims contradicted by the media.

A joint verifier receives the observation deltas and checks audio, visual, cross-modal, and agentic consistency against the media and trajectory state available at the current step. For agentic verification, candidate findings identify a specific decision, the relevant evidence, and the suspected error. The verifier confirms, rejects, or revises each candidate. The criteria cover failure to revise contradicted hypotheses, repeated searches without new information, missing comparisons, inappropriate tool selection, and premature or delayed stopping.

D.3 Diagnosis-Guided Refinement

At the selected step, we restore the original model-visible history and media. The failed turn is supplied as an unexecuted draft; subsequent actions and observations from the original trajectory are excluded. The teacher generates an updated observation, a concise draft preserving the original reasoning error, a correction, and a new action. The draft and correction are combined into the reasoning field of the standard interaction format, and only the new action is executed. Private diagnostic guidance is removed from the exported conversation.

The judge does not receive the reference answer during diagnosis or recovery. A candidate is retained only when its final answer passes the same outcome check as the source trajectory. Attempts restart from the same source state with different random seeds, stopping at the first accepted candidate or after max attempts.

E Evidence Reward Details

E.1 Modality Matching and Task-Specific Scoring

For task x, the score $q \in [0, 1]$ compares a selected subset of retrieved evidence $\mathcal{B}$ with the task's evidence annotations $\mathcal{A}_x$. The reward schema distinguishes visual, audio, and joint audio-visual evidence, denoted by V, A, and AV. A candidate's requested modality determines which evidence it can support; a joint audio-visual candidate may support visual, audio, or joint audio-visual targets when the target annotation permits that match. Let $\mathcal{M}_x$ contain the annotated modalities with positive modality weights λ_m. For modality $m \in \mathcal{M}_x$, write the target annotations as $\mathcal{A}_x^m = \{(\tau_j, w_j)\}_{j=1}^{J_m}$, where τ_j is a temporal interval, $w_j > 0$ its importance weight, and J_m the number of targets. Let P_m be the union of intervals in the selected subset $\mathcal{B}$ with modality m. The

combined score is

$$q(\mathcal{B}, \mathcal{A}_x) = \frac{\sum_{m \in \mathcal{M}_x} \lambda_m q_m(P_m, \mathcal{A}_x^m)}{\sum_{m \in \mathcal{M}_x} \lambda_m}, \tag{10}$$

where q_m is the task-specific score below. An annotated modality with no selected candidate receives zero; modalities absent from the targets do not contribute. The scorer defaults to equal modality weights and unit target weights. Overlapping intervals are merged before computing support duration, denoted by $|\cdot|$.

Question answering. QA annotations identify supporting evidence without exhaustively labeling all useful context. For a fixed modality, define the target union $G_m = \bigcup_{j=1}^{J_m} \tau_j$, total weight $Z_m = \sum_{j=1}^{J_m} w_j$, and weighted support duration $D_m = \sum_{j=1}^{J_m} w_j |\tau_j|$. Coverage c_m and locality ℓ_m are

$$\begin{aligned} c_m &= \frac{1}{Z_m} \sum_{j=1}^{J_m} w_j \frac{|P_m \cap \tau_j|}{|\tau_j|}, \\ \ell_m &= \begin{cases} \left(1 + \dfrac{|P_m \setminus G_m|}{D_m}\right)^{-1}, & |P_m \cap G_m| > 0, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{11}$$

The QA score is $q_m = \alpha c_m + (1 - \alpha)\ell_m$. Coverage rewards retrieval of annotated support, while locality favors a more focused interval once support has been reached. The scorer uses $\alpha = 0.5$ by default.

Temporal grounding. TG targets specify the interval set to be localized. For non-empty temporal interval unions P and G, let $\phi(P, G)$ denote the grounding reward for valid predictions in TimeLens2 (Zhu et al., 2026), scaled by $1/2$ to lie in $[0, 1]$. For equal target weights, $q_m = \phi(P_m, G_m)$. For unequal weights, the scorer uses $q_m = Z_m^{-1} \sum_{j=1}^{J_m} w_j \phi(P_m, \tau_j)$. An empty retrieved support receives zero.

E.2 Selected Evidence State and Incremental Credit

For task x, let $\mathcal{C}_t$ denote the collection of retrieved intervals and their modalities available after step t. The ideal selected-evidence utility is

$$U_t^\star = \max_{\mathcal{B} \subseteq \mathcal{C}_t} q(\mathcal{B}, \mathcal{A}_x). \tag{12}$$

The evidence scorer starts with $\mathcal{C}_0 = \emptyset$ and $U_0 = 0$. Subsequent collections accumulate the temporal supports and modalities returned by successful tool calls. The scorer retains this interval metadata even when earlier raw media leave the model's active context. The agent receives an initial overview, but this overview is not added to the evidence candidate collection used for scoring.

The subset maximization separates by modality because q is a weighted sum of independent modality scores. The implementation enumerates all subsets when a modality has at most 12 candidates, preferring fewer candidates when scores tie. Above this limit, it ranks candidates by their singleton score, retains the top 12, and enumerates subsets of this reduced collection. Equal singleton scores are ordered by candidate identifier.

Let $\widetilde{U}_t$ be the utility returned by this bounded search at step t. Because the reduced candidate set can change, the implemented state retains the previous utility whenever a new approximation decreases it:

$$U_0 = \widetilde{U}_0, \qquad U_t = \max\{U_{t-1}, \widetilde{U}_t\}, \qquad \Delta U_t = U_t - U_{t-1}. \tag{13}$$

For trajectory i and turn k, the corresponding quantities are written as $U_{i,k}$ and $\Delta U_{i,k}$ in Sec. 3.2. Exact and approximate search modes are recorded separately. With exact search, $U_t = U_t^\star$. The retained utility is non-decreasing, and the non-discounted increments sum to $U_T - U_0$ over T steps in both search modes.

E.3 Step-Level Rewards and Policy Optimization

Trajectory reward. The answer reward R_i^{ans} depends on the task type. Multiple-choice answers receive 1 for a matching option label and 0 otherwise. Free-form answers first undergo normalized exact matching; unmatched answers are judged against the question and reference answer, receiving 1 if correct and 0 otherwise. For temporal grounding, the answer reward is $\phi(P, G)$ defined in Appendix E.1, where P and G are the predicted and reference interval unions. An absent or invalid final answer receives zero answer reward. The protocol reward R_i^{fmt} is $+0.1$ when the final answer uses the canonical format and the trajectory contains no invalid or no-progress actions, and -0.1 otherwise. We sample $G = 8$ trajectories per instruction and subtract their mean total reward to obtain A_i, without dividing by the group standard deviation.

$$R_i = R_i^{\text{ans}} + R_i^{\text{fmt}}, \qquad A_i = R_i - \frac{1}{G}\sum_{j=1}^{G} R_j. \tag{14}$$

Process score. The process score $p_{i,k}$ in Sec. 3.2 combines the incremental evidence utility with protocol indicators:

$$\begin{aligned} p_{i,k} &= 0.2\,\Delta U_{i,k} - 0.1\,I_{i,k}^{\text{invalid}} - 0.05\,I_{i,k}^{\text{repeat}}, \\ z_{i,k} &= \frac{p_{i,k} - \mu_i}{\sigma_i + \epsilon}. \end{aligned} \tag{15}$$

The binary indicators mark invalid and repeated calls, matching the definitions in the main text. The statistics μ_i and σ_i are the mean and population standard deviation of $p_{i,k}$ across real generated turns in trajectory i. We use $\epsilon = 10^{-6}$ and $\lambda = 0.1$. For tasks without evidence annotations, the evidence term is zero, while the protocol penalties remain active. These indicators use the interaction protocol's action checks and duplicate-request detection. A retrieval with a different modality or precision is not penalized merely because its temporal interval overlaps a previous request. The process score modulates turn advantages; it is not added to the trajectory reward or used for candidate filtering.

Within-trajectory adjustment. Let K_i be the number of real turns in trajectory i, including the final-answer turn. The normalization statistics are

$$\mu_i = \frac{1}{K_i}\sum_{k=1}^{K_i} p_{i,k}, \qquad \sigma_i = \sqrt{\frac{1}{K_i}\sum_{k=1}^{K_i}(p_{i,k} - \mu_i)^2}. \tag{16}$$

Equivalently, each turn advantage is $\widetilde{A}_{i,k} = A_i w_{i,k}$ with

$$w_{i,k} = 1 + \lambda\,\text{clip}(\text{sign}(A_i) z_{i,k}, -1, 1) \in [0.9, 1.1]. \tag{17}$$

A single-turn trajectory or constant process scores give $w_{i,k} = 1$. A zero trajectory advantage remains zero after the adjustment. Clipping bounds the change to each turn's advantage but does not enforce an exactly conserved sum of advantages across turns.

Policy update. Each real turn is passed to the trainer as a separate response, and its generated tokens inherit $\widetilde{A}_{i,k}$. The importance ratio is the exponential of the mean token log-probability difference between the current and rollout policies within that turn. The clipped objective uses lower and upper ratio bounds of 0.8 and 1.28. Response-token losses are summed and normalized by the total number of trainable response tokens in the trajectory. We add a KL loss against a fixed reference policy with coefficient 0.001. Invalid and no-progress action tokens remain trainable, while operational failures are excluded from the policy loss.

F Limitations and Future Work

Our study is limited to a few omni-video backbones and benchmarks, while trajectory construction relies on strong teacher models and multi-stage verification, incurring additional cost. Reflection trajectories improve error recovery but do not cover all possible failure modes. Future work will explore (1) more efficient data construction within the data model in the loop, (2) stronger self-improvement mechanisms for omni-video agents, including enhance our stage 3 design in the data learning agent loop.