RE-OMNI: A TWO-PASS REINFORCEMENT LEARNING RECIPE FOR AGENTIC VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Agentic video understanding commonly uses multi-turn loops to alternate reasoning and evidence acquisition. However, multi-turn configurations generate thousands of output tokens yet underperform the instruction-tuned supervised fine-tuning (SFT) baseline based on question answering (QA) data for limited model scales. To address these problems, we introduce Re-Omni, a two-pass reinforcement learning (RL) recipe enabling direct training from a strong instruction-tuned model without teacher-generated chain-of-thought (CoT) annotation or a separate CoT cold-start SFT stage. Re-Omni combines three mechanisms: **re-watch** revisits a predicted relevant interval at higher spatial fidelity; **re-ask** repeats the question after the new evidence; and **re-answer** refines an initial direct answer. Re-Omni improves accuracy over QA-SFT on all six multiple-choice video benchmarks. Against the full-history multi-turn configuration, it gains more than 4.1 percentage points across the benchmarks with two viewing rounds, reducing generated output tokens from 2,681.2 to 101.7 and inference GPU time from 18.3 to 9.5 seconds per question. Code, models, and data will be publicly released.

1 INTRODUCTION

Video large language models (video-LLMs) must balance access to fine-grained evidence against the computational cost of processing long, high-resolution videos. Sparse frame sampling and token compression make video understanding tractable, but can discard details needed to answer a specific question (Liu et al., 2025b; Li et al., 2025; Qin et al., 2025; Sun et al., 2025). Agentic video understanding addresses this limitation by allowing models to actively localize and revisit relevant video segments. This shifts evidence acquisition from a fixed sampling procedure to a question-dependent process guided by the model. The key is to choose where a second, more detailed view can most improve the final answer.

Recent work largely implements evidence acquisition through multi-turn loops that interleave reasoning with actions. At the system level, VideoLucy and GCAgent coordinate specialized agents to retrieve content through textual memories (Zuo et al., 2025; Yeo et al., 2025). Within a single model, LOVE-R1, LongVT, and VideoZoomer use repeated fine-grained inspection or temporal localization (Fu et al., 2026; Yang et al., 2026; Ding et al., 2026). These approaches couple adaptive viewing with repeated reasoning. Many single-model methods also use a chain-of-thought (CoT) cold-start stage before reinforcement learning (RL), adding reasoning supervision to the training recipe.

However, in our 8B-model comparison, both multi-turn configurations underperform the instruction-tuned QA-SFT baseline on most benchmarks while generating thousands of output tokens per question. Configurations with CoT cold-start also remain below the direct-RL alternative on most reported endpoints. These observations, together with prior findings that CoT-based SFT does not consistently improve video understanding Feng et al. (2025), motivate asking whether targeted re-watching can improve answers without a long reasoning trajectory or an additional CoT cold-start stage.

We introduce Re-Omni, a fixed two-pass policy trained with DAPO (Yu et al., 2025) from the audio-visual caption-SFT initialization. It gives an initial answer, selects an interval, then re-watches and refines the answer. Answer-level and format rewards train this behavior without teacher-generated CoT traces, ground-truth timestamps, or a separate CoT cold-start SFT stage.

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.

Three complementary mechanisms define the workflow. **Re-watch** predicts a query-relevant interval after an initial global viewing and revisits it with a higher spatial-resolution budget. **Re-ask** repeats the question after the revisited evidence. Under causal attention, the original question cannot incorporate later video tokens, whereas the repeated question can attend to them; the final answer can then use both. **Re-answer** places an initial direct answer before localization and refines it after re-watching, retaining the instruction-tuned model's familiar answer-first prefix.

Across six benchmarks, the primary matched-data-and-backbone comparison shows gains over the baseline. The workflow study further shows higher endpoint accuracy than the evaluated multi-turn configurations, with fewer viewing rounds and substantially fewer generated output tokens. Our contributions are summarized as follows:

- We show that, in our 8B-model comparison, the evaluated multi-turn configurations generate much longer outputs but underperform the instruction-tuned baseline on most benchmarks.
- We introduce Re-Omni, a two-pass re-watch, re-ask, and re-answer policy trained directly using RL alone, without teacher-generated CoT annotation or a separate CoT cold-start SFT stage.
- We demonstrate gains over the same-data baseline on all six benchmarks and use separately trained ablations to evaluate question repetition, targeted re-watching, and answer refinement.

2 Related Works

2.1 Multimodal LLMs for Video Understanding

Video-LLMs typically connect a visual encoder to an LLM through a lightweight projector and represent videos as sparsely sampled frames. Examples include LLaVA-Video, VideoLLaMA 3, Qwen3-VL, and Intern-VL 3.5 (Zhang et al., 2024c; 2025; Bai et al., 2025; Wang et al., 2025a). Audio-visual models further incorporate acoustic evidence for joint reasoning, including VideoLLaMA 2, video-SALMONN 1 and 2, Qwen2.5-Omni, Qwen3-Omni, AVoCaDO, Omni-Captioner, D-ORCA, and MiniCPM-o 4.5 (Cheng et al., 2024; Sun et al., 2024; Tang et al., 2025; Xu et al., 2025a;b; Chen et al., 2026; Ma et al., 2026; Tang et al., 2026; Cui et al., 2026).

Memory and compute budgets limit temporal and spatial fidelity, which can obscure the evidence needed for a specific question. Prior methods reduce visual tokens through multi-frame fusion (Liu et al., 2025b; Fei et al., 2026; Li et al., 2025), chunk-wise prefilling for long videos (Qin et al., 2025; Wang et al., 2025b; Gao et al., 2026), and training-free token compression (Liu et al., 2025a; Zhang et al., 2024b; Kim et al., 2025). Streaming methods maintain compact memories of earlier video content (Qian et al., 2024; Sun et al., 2025; Zhang et al., 2024a; Qian et al., 2025). These approaches compress or retain information according to their respective representations and selection rules. Re-watching takes a complementary route: it lets the model select evidence across viewing passes based on the question and its first view of the video.

2.2 Agentic Video Understanding through Re-Watching

Temporal localization lets video models revisit informative segments without processing every frame at high fidelity. Agentic systems can coordinate specialized models to retrieve evidence or let one video model decide where to look next. The latter design is closest to Re-Omni; within this single-model family, methods differ in their interaction depth, answer generation, and training supervision.

Multi-agent retrieval. VideoLucy (Zuo et al., 2025) organizes textual video memories hierarchically and uses agent-based backtracking to retrieve finer details. GCAgent (Yeo et al., 2025) couples a memory manager with a reasoning agent in a perception–action–reflection loop. Both acquire evidence through repeated coordination among specialized components and textual memories.

Single-model interaction depth. Recent methods integrate temporal localization and evidence acquisition into one compact video model, following the recipe of Group Relative Policy Optimization (GRPO) (Guo et al., 2025) and Reason and Act (ReAct) (Yao et al., 2023). LOVE-R1, LongVT, and VideoZoomer each use a single 7B-scale video model to select and inspect clips (Fu et al., 2026; Yang et al., 2026; Ding et al., 2026). LOVE-R1 permits repeated zoom-ins, LongVT interleaves reasoning with video-cropping calls, and VideoZoomer performs an initial glance followed by repeated temporal

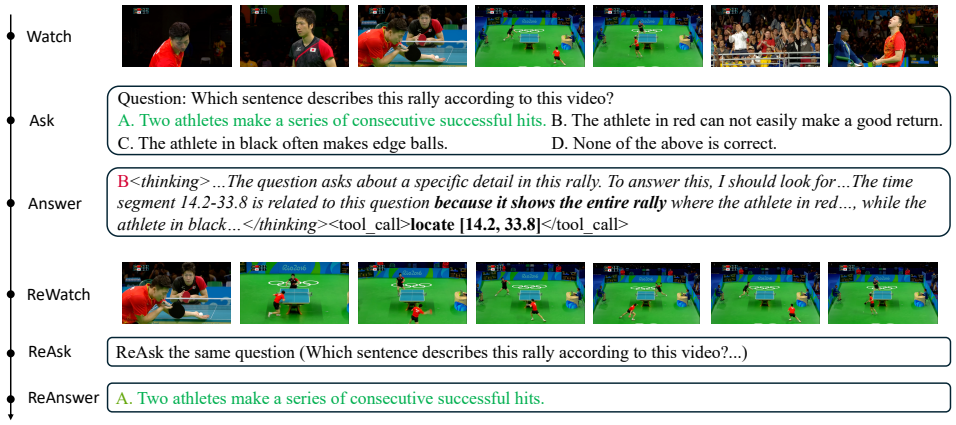

Figure 1: Overview of the Re-Omni workflow. Given a video and question, the model first **watches** the full video under frame and spatial-resolution caps, **asks** the question, and produces an initial **answer**, followed by a reasoning trace and temporal localization. It then **re-watches** the selected interval at higher spatial fidelity, **re-asks** the question, and **re-answers** by retaining or revising the initial answer. The second pass keeps the global first-pass context while adding frames from the selected interval.

zooming. In each case, the same model can request further evidence over multiple rounds. Re-Omni also uses one 8B model, but fixes the interaction to a global view followed by one targeted re-watch.

Second-pass reasoning. Reducing viewing rounds does not itself eliminate repeated reasoning. Re-Omni first gives a direct answer, then reasons about where to look, and finally answers again after re-watching without a second explicit reasoning trace. Reasoning guides evidence acquisition once; re-ask and re-answer then connect the revisited evidence to the final response. This answer-first localization also preserves the instruction-tuned model's question-to-answer prefix.

Training supervision. LOVE-R1, LongVT, and VideoZoomer use supervised cold-start stages before RL to teach reasoning-and-action behavior (Fu et al., 2026; Yang et al., 2026; Ding et al., 2026). Re-Omni directly optimizes its two-pass policy from a strong instruction-tuned baseline, without teacher-generated CoT annotations, ground-truth temporal localization annotations, or an additional CoT-SFT cold-start stage. The combined difference is a fixed two-pass workflow with direct second-pass answering and no agentic cold-start.

3 METHODS

3.1 MODEL WORKFLOW

The workflow of Re-Omni is shown in Fig. 1. We interleave visual and audio tokens produced by relevant encoders to form a unified video representation $\mathbf{V} = \mathrm{Interleave}(\mathbf{V}_{\text{video}}, \mathbf{V}_{\text{audio}})$, which, together with the question $\mathbf{Q}$, is fed into the LLM to produce an output $\mathbf{O}$:

$$\mathbf{O} = \mathrm{LLM}(\mathbf{V}, \mathbf{Q}). \tag{1}$$

First pass (watch). The model processes the full video under frame and spatial-resolution caps to obtain a coarse global representation $\mathbf{V}^{(1)} = \mathrm{Interleave}(\mathbf{V}^{(1)}_{\text{video}}, \mathbf{V}^{(1)}_{\text{audio}})$. Conditioned on $\mathbf{V}^{(1)}$ and $\mathbf{Q}$, the LLM produces a structured output $\mathbf{O}^{(1)}$ comprising three components: an initial answer $\mathbf{A}^{(1)}$, a short reasoning trace $\mathbf{R}^{(1)}$, and a temporal localization prediction $\mathbf{T} = [t_{\text{start}}, t_{\text{end}}]$ identifying the segment most relevant to the query:

$$\mathbf{O}^{(1)} = (\mathbf{A}^{(1)}, \mathbf{R}^{(1)}, \mathbf{T}) = \mathrm{LLM}(\mathbf{V}^{(1)}, \mathbf{Q}). \tag{2}$$

Here, $\mathbf{A}^{(1)}$ leverages the well-aligned prior of the instruction-tuned base model, while $\mathbf{T}$ determines where higher-fidelity processing will be applied in the next pass.

Second pass (re-watch). Guided by $\mathbf{T}$, the model revisits the localized segment with a higher spatial-resolution budget, yielding a fine-grained representation $\mathbf{V}^{(2)} = \mathrm{Interleave}(\mathbf{V}^{(2)}_{\text{video}}, \mathbf{V}^{(2)}_{\text{audio}})$ sampled from $[t_{\text{start}}, t_{\text{end}}]$. During training, we reject intervals whose duration would make the second-pass effective frame rate lower than in the first pass; when the first pass is below 10 FPS, the second pass must be strictly denser. This ensures that accepted re-watches preserve or increase temporal detail while using a higher spatial-resolution budget. To integrate global context and local detail, we retain the first-pass input $\mathbf{V}^{(1)}$ and the interaction trace $(\mathbf{Q}, \mathbf{O}^{(1)})$. During the second generation call, we directly reuse the entire KV cache for this first-pass prefix, avoiding recomputation of its keys and values and reducing inference cost. Under causal attention, the later video tokens $\mathbf{V}^{(2)}$ can attend to the original question, whereas the original question representations cannot incorporate $\mathbf{V}^{(2)}$. We re-inject the question as $\mathbf{Q}'$ after $\mathbf{V}^{(2)}$; structurally, $\mathbf{Q}'$ can attend to $\mathbf{V}^{(2)}$, and the subsequent answer tokens can attend to both the original question and the later video. The refined answer is produced as:

$$\mathbf{O}^{(2)} = \mathbf{A}^{(2)} = \mathrm{LLM}\left(\mathbf{V}^{(1)}, \mathbf{Q}, \mathbf{O}^{(1)}, \mathbf{V}^{(2)}, \mathbf{Q}'\right). \tag{3}$$

The final prediction $\mathbf{A}^{(2)}$ either retains $\mathbf{A}^{(1)}$ when the initial evidence suffices or revises it when higher-fidelity inspection reveals additional details. This two-pass watch–re-watch loop forms the core computational unit of Re-Omni and serves as the trajectory for subsequent RL.

3.2 Training Procedure

Training proceeds in three stages: **1)** audio alignment, **2)** audio-visual caption SFT, and **3)** end-to-end RL. The first two stages establish a strong instruction-tuned multimodal base model, while the final stage injects re-watch behavior without CoT-based cold-start. Unless otherwise specified, we treat audio as an integral component of the video modality throughout this paper.

Audio alignment. We first augment a visual-only LLM with an audio branch by training an audio projector that maps acoustic features into the LLM representation space (Tang et al., 2024). During this stage, only the projector is updated, while the audio encoder, visual encoder, and LLM backbone remain frozen. Training uses large-scale speech recognition and audio captioning data under a standard cross-entropy objective. This step establishes an audio interface without updating the existing visual branch or LLM backbone.

Audio-visual caption SFT. Next, we jointly model audio-visual inputs by fine-tuning on densely annotated video captions (Tang et al., 2025). The model is trained to generate unified textual descriptions of video content from interleaved audio and visual tokens. We employ low-rank adaptation (LoRA) (Hu et al., 2022) on the LLM backbone and jointly optimize it with the modality projectors, while keeping the audio and visual encoders frozen. This stage produces the instruction-tuned base video-LLM used in our experiments.

End-to-end RL. RL optimizes the complete two-pass trajectory by updating only LoRA adapters; all base weights and other modules remain frozen. We inject re-watch capability without relying on CoT cold-start SFT. Instead of constructing large-scale annotated reasoning trajectories, we define a lightweight system prompt specifying the interaction protocol and the `<thinking>` / `<tool_call>` format, and optimize full trajectories with rule-based rewards using the DAPO objective in Eqn. (6).

Under the **re-answer** design, $\mathbf{A}^{(1)}$ appears immediately after $(\mathbf{V}^{(1)}, \mathbf{Q})$, with no reasoning scaffold or tool interaction inserted before it. This preserves the base model's question-to-answer prefix interface and gives RL an answer-first trajectory structure. Additional implementation details are provided in Appendix B.

Reward design. The trajectory reward is defined as a weighted sum of five rule-based components, each taking values in $\{0, 1\}$:

- r_{acc1}, r_{acc2}: are accuracy rewards indicating whether the initial answer $\mathbf{A}^{(1)}$ and the refined answer $\mathbf{A}^{(2)}$ match the ground-truth option, respectively, encouraging correct answers in both passes.
- r_{fmt1}: is a format reward indicating whether the first pass follows the "answer–think–locate" schema with a parseable tool call that selects a valid, sufficiently short interval.

- r_{fmt2}: is a format reward indicating whether the second pass contains no meta-tokens (*e.g.*, <thinking> for thinking and <tool_call> for tool calling), enforcing a clean final answer.
- r_{rev}: for trajectories reaching the second pass, is a revise reward that equals 0 only when $\mathbf{A}^{(1)}$ is incorrect and $\mathbf{A}^{(2)}$ simply repeats it, and 1 otherwise, encouraging the model to revise incorrect initial predictions.

During training, a valid interval must satisfy $0 \leq t_{\text{start}} < t_{\text{end}} \leq T$ for video duration T, select less than the full video, and meet the frame-rate condition specified below. A call that covers the full video or requests an excessively long interval is invalid. If the first-pass tool call is invalid or unparseable, the trajectory terminates without re-watching or generating a second answer. Such a trajectory is counted as an incorrect final prediction, with $r_{\text{fmt1}} = 0$ and all second-pass rewards set to zero: $r_{\text{acc2}} = r_{\text{fmt2}} = r_{\text{rev}} = 0$. The total reward for the i th trajectory $\mathbf{o}_i$ is computed as

$$\mathcal{R}(\mathbf{o}_i) = \lambda_{\text{acc1}} r_{\text{acc1}}(\mathbf{o}_i) + \lambda_{\text{acc2}} r_{\text{acc2}}(\mathbf{o}_i) + \lambda_{\text{fmt1}} r_{\text{fmt1}}(\mathbf{o}_i) + \lambda_{\text{fmt2}} r_{\text{fmt2}}(\mathbf{o}_i) + \lambda_{\text{rev}} r_{\text{rev}}(\mathbf{o}_i), \quad (4)$$

where the λs are coefficients that balance the contributions of each component. The group-normalized advantage for each token is then defined as

$$\hat{\mathcal{A}}_{i,t} = \frac{\mathcal{R}(\mathbf{o}_i) - \text{mean}(\mathcal{R})}{\text{std}(\mathcal{R}) + \epsilon}, \quad (5)$$

where ϵ is a small constant for numerical stability, and $\text{mean}(\cdot)$ and $\text{std}(\cdot)$ denote the mean and standard deviation. Collectively, these terms reward answer correctness and structural validity and penalize an unchanged second-pass answer when the initial answer is incorrect.

RL objective. We use the token-level DAPO objective (Yu et al., 2025). For each question q, the rollout policy $\pi_{\theta_{\text{old}}}$ generates G trajectories $\{\mathbf{o}_i\}_{i=1}^{G}$. The policy is updated by minimizing the clipped surrogate loss

$$\mathcal{J}(\theta) = -\mathbb{E}_{q,\{\mathbf{o}_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}} \left[\frac{1}{\sum_{i=1}^{G} |\mathbf{o}_i|} \sum_{i=1}^{G} \sum_{t=1}^{|\mathbf{o}_i|} \min\Big(\rho_{i,t}(\theta) \hat{\mathcal{A}}_{i,t},\ \text{clip}\big(\rho_{i,t}(\theta), 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}\big) \hat{\mathcal{A}}_{i,t} \Big) \right], \quad (6)$$

where $\rho_{i,t}(\theta) = \pi_\theta(\mathbf{o}_{i,t} \mid q, \mathbf{o}_{i,<t}) / \pi_{\theta_{\text{old}}}(\mathbf{o}_{i,t} \mid q, \mathbf{o}_{i,<t})$, $\hat{\mathcal{A}}_{i,t}$ is defined in Eqn. (5), and ϵ_{low} and ϵ_{high} set the lower and upper clipping ranges. Normalizing by the total number of generated tokens gives each token equal weight in the loss.

4 Experiment Setup

4.1 Model Specifications

Re-Omni is built on Qwen3-VL with 8 billion (B) parameters (Bai et al., 2025). Its audio features are extracted by the Whisper-Large-v3 encoder (Radford et al., 2023) and subsequently aligned through a window-level Q-Former with a 0.5-second window.

In the first pass, visual frames are sampled at up to 10 frames per second (FPS) with a maximum of 768 frames. Videos longer than 76.8 seconds are uniformly sub-sampled to meet this budget, and each frame is capped at 44,100 pixels to maintain a broad but low-fidelity global view under a constrained token budget. In the second pass, the localized segment is sampled at up to 10 FPS with a limit of 128 frames. Each frame is allowed up to 176,400 pixels to preserve fine-grained spatial details within the selected interval. For a video of duration T and a proposed interval of duration $\Delta t = t_{\text{end}} - t_{\text{start}}$, the nominal effective rates under these frame caps are $f_1 = \min(10, 768/T)$ and $f_2 = \min(10, 128/\Delta t)$ FPS. During training, we accept the interval only if $f_2 > f_1$ when $f_1 < 10$ FPS, or $f_2 = f_1 = 10$ FPS when the first pass already reaches the sampling-rate limit. For instance, an interval spanning the full video will be rejected. Thus, a proposed re-watch must provide at least as much temporal detail as the first pass, with strictly more detail when the first pass was sub-sampled. This duration check prevents a long localization call from merely reproducing a sparse global view. Although this configuration can theoretically produce up to approximately 27,000 video tokens across both passes, the actual count is further regulated by the visual pre-processor's smart resizing policy, which keeps the total number of video tokens around 24,000 per trajectory in practice.

4.2 Training Specifications

For the audio alignment stage, we train only the audio aligner while keeping the audio encoder, the visual branch, and the LLM backbone frozen, using a combination of LibriSpeech 960-hour (Panayotov et al., 2015) and CommonVoice (Ardila et al., 2020) for automatic speech recognition, together with WavCaps (Mei et al., 2024) and AudioCaps (Kim et al., 2019) for audio captioning, to map acoustic features into the representation space of the LLM. For the audio-visual caption SFT stage, we fine-tune the model on audio-visual captions derived from LLaVA-Video-178k (Zhang et al., 2024c) and re-annotated with Gemini 2.5 Pro (Comanici et al., 2025), which jointly describe the visual and acoustic content within a unified temporal narrative and produce the instruction-tuned audio-visual base model used in our experiments.

For the end-to-end RL stage, we use CinePile (Rawal et al., 2024) and CG-Bench (Chen et al., 2025) as training sources for short- and long-form videos, respectively, covering a broad range of temporal scales. To obtain a more informative difficulty distribution, we perform 8 rollouts using the audio-visual caption SFT model over the entire training set and discard samples that are answered either correctly or incorrectly in all attempts, retaining approximately 110k moderately challenging samples. We optimize with Adam for 2,433 updates. The learning rate warms up linearly over the first 3% of updates to 5×10^{-6} and then follows cosine decay. We use a group size of 8 and reward coefficients $\lambda_{\text{acc1}} = 0.9$, $\lambda_{\text{acc2}} = 1.1$, $\lambda_{\text{fmt1}} = 1$, $\lambda_{\text{fmt2}} = 1$, and $\lambda_{\text{rev}} = 0.5$, slightly emphasizing the refined answer during optimization. Rollouts use temperature 0.95, top-k 20, and top-p 0.95. We report the final checkpoint rather than selecting a checkpoint on the evaluation sets. Reward curves are provided in Appendix C, and additional ablations on the RL settings can be found in Appendix D.

4.3 Evaluation Specifications

We evaluate Re-Omni on six multiple-choice benchmarks spanning short-to-medium and long videos. For short-to-medium videos, we use *VideoHolmes* (Cheng et al., 2025), *DailyOmni* (Zhou et al., 2025), *AVUT* (Yang et al., 2025), and *OmniVideoBench* (Li et al., 2026), which assess fine-grained audio-visual reasoning within relatively tight temporal windows. For long videos, we use *VideoMME* (Fu et al., 2025) and *LVOmniBench* (Tao et al., 2026), which assess understanding over videos spanning tens of minutes to hours. The six test sets contain 9,482 unique QA pairs in aggregate. Duplicate videos may occur within the training collection and within the benchmark collections, but a video-level comparison found no video shared between training and evaluation. All results use the same evaluation implementation and greedy decoding. An output that cannot be parsed as a valid answer choice is scored as incorrect. For the primary QA-SFT/Re-Omni comparison, we use three paired runs with seeds 2024, 2025, and 2026 and report the mean and sample standard deviation. Other configurations and ablations are single-run results: Tables 2–4 report the currently evaluated Re-Omni checkpoint rather than the three-run mean.

5 Experiments

5.1 Overall Results

Table 1 compares Re-Omni with two baselines built on the same backbone: AV-Caption-Base and QA-SFT. QA-SFT shares Re-Omni's caption pre-training, QA training data, batch size, learning rate, optimizer, and number of updates, but uses supervised learning and single-pass answering instead of the RL-trained two-pass workflow. QA-SFT improves on AV-Caption-Base on five benchmarks, although its AVUT accuracy drops from 76.7 to 74.1. Re-Omni improves accuracy over QA-SFT by mean paired gains of 1.1, 0.5, 3.4, 2.9, 2.9, and 2.4 percentage points across all six benchmarks. The largest gains occur on AVUT, OmniVideoBench, and VideoMME; on AVUT, Re-Omni also exceeds AV-Caption-Base, recovering the loss seen after QA-SFT. Gains of 2.9 and 2.4 points on the two long-video benchmarks show that the benefit extends beyond the short-to-medium-video tests. All 18 matched seed–benchmark comparisons favor Re-Omni. Five of the six 95% intervals for the mean paired gains exclude zero. Appendix F provides the individual runs and interval calculations.

We also show reported scores for Qwen 2.5-Omni (Xu et al., 2025a), video-SALMONN 2+ (Tang et al., 2025), Qwen 3-Omni (Xu et al., 2025b), D-ORCA (Tang et al., 2026), and Qwen 3-VL (Bai et al., 2025). Re-Omni has the highest listed mean in five of six columns, exceeding the strongest

Table 1: Accuracy on six short-to-medium and long-form benchmarks. The primary matched-data-and-backbone comparison reports mean ± sample standard deviation over three paired runs (seeds 2024, 2025, and 2026). External-model rows are contextual reported point estimates without fully matched evaluation protocols. Bold denotes the highest numerical accuracy in each column.

Model	Short-to-Medium				Long	
	VideoHolmes	DailyOmni	AVUT	OmniVideoBench	VideoMME	LVOmniBench
Qwen 2.5-Omni	43.7	62.7	66.3	29.3	64.3	32.0
video-SALMONN 2+	46.9	71.8	69.5	36.4	73.4	32.7
Qwen 3-Omni	54.1	69.8	72.0	38.4	70.5	35.8
D-ORCA	48.5	**78.5**	76.1	-	72.9	-
Qwen 3-VL	46.6	60.1	61.4	-	71.4	35.6
AV-Caption-Base	49.8	76.7	76.7	40.0	72.5	39.3
QA-SFT	53.4±0.2	77.9±0.0	74.1±0.1	40.8±0.0	72.8±0.1	40.5±0.1
Re-Omni	**54.5**±0.1	78.4±0.3	**77.5**±0.1	**43.7**±0.4	**75.8**±0.6	**42.9**±0.1

reported external score by 0.4 points on VideoHolmes, 1.4 on AVUT, 5.3 on OmniVideoBench, 2.4 on VideoMME, and 7.1 on LVOmniBench. The largest margins occur on OmniVideoBench and LVOmniBench, indicating strong performance on both fine-grained short-to-medium video understanding and long-form video understanding. Together with the same-backbone comparisons, these results support the effectiveness of Re-Omni's RL-trained two-pass workflow across the evaluated video understanding tasks.

5.2 Progressive Simplification of Agentic Video Understanding

Table 2 compares nine recipes, including QA-SFT as a same-data supervised reference and Base (AV-Caption-Base before task-specific QA training). Both multi-turn variants, Two-turn, and Reasoning once use approximately 5,000 Gemini-annotated CoT cold-start examples (Comanici et al., 2025), then the same RL stage as Re-Omni; Direct RL and the final Re-Omni skip cold-start. Full history retains previous visual inputs and reasoning; latest video replaces earlier visual inputs with reasoning summaries, allowing more video tokens per round. Two-turn uses a global view and targeted revisit, retains first-pass context, and reasons in both turns; Reasoning once preserves this protocol but answers directly after revisiting. Direct RL starts from instruction-tuned Base with re-watch alone, without a repeated question or initial answer to refine; Re-Omni adds re-ask and answer-first re-answer. The two multi-turn variants serve as approximations of LOVE-R1, LongVT, and VideoZoomer (Fu et al., 2026; Yang et al., 2026; Ding et al., 2026), reproducing their repeated reasoning-and-re-watching pattern on our shared backbone and evaluation setup rather than their full model-specific implementations.

Table 2 shows that progressively simplifying the evaluated workflows improves mean accuracy across the four benchmarks while reducing inference cost. The two multi-turn variants achieve arithmetic mean accuracies of 58.8% (full history) and 58.5% (latest video), with 2,681.2–4,112.9 generated output tokens and 18.3–24.0 s of GPU time per question. Restricting the workflow to Two-turn raises mean accuracy to 60.1%; removing second-pass reasoning further increases it to 60.3% with Reasoning once, while output length falls to 1,134.0 and 734.2 tokens and GPU time to 12.5 and 11.0 s, respectively. Direct RL, which also skips CoT cold-start, reaches 60.7% with 109.2 output tokens and 9.6 s of GPU time. Adding re-ask and answer-first re-answer yields Re-Omni's 63.1% mean accuracy, a further gain of 2.4 percentage points at a similar cost of 101.7 output tokens and 9.5 s. Overall, Re-Omni improves mean accuracy by 4.4–4.7 points over the multi-turn variants while reducing GPU time by 48%–60%. Relative to QA-SFT's 61.2% mean accuracy, Re-Omni gains 1.9 points, with GPU time increasing from 6.1 to 9.5 s.

To isolate the effect of cold-start from the other recipe changes, we also evaluate Re-Omni + cold-start, which yields lower accuracy and higher inference cost than Re-Omni.

Table 2: Progressive recipe study. Accuracy (%) is reported on VideoHolmes (VH), DailyOmni (DO), VideoMME (VMME), and LVOmniBench (LVO), followed by average input and generated-output tokens and inference GPU time per question. Base denotes AV-Caption-Base. The two multi-turn rows approximate the repeated reasoning-and-re-watching workflow of the cited multi-turn models.

Configuration	**VH**	**DO**	**VMME**	**LVO**	**Avg. input**	**Avg. output**	**Avg. GPU time (s)**
Base	49.8	76.7	72.5	39.3	26031.4	4.2	6.1
QA-SFT	53.4	77.9	72.9	40.6	26031.4	4.1	6.1
Multi-turn (full history)	49.6	74.6	72.2	38.6	37751.2	2681.2	18.3
Multi-turn (latest video)	50.5	74.4	71.9	37.1	23739.7*	4112.9	24.0
Two-turn	51.8	76.5	73.4	38.8	24592.6	1134.0	12.5
Reasoning once	52.5	77.0	73.0	38.6	24081.4	734.2	11.0
Direct RL	53.1	77.6	72.9	39.2	23994.8	109.2	9.6
Re-Omni + cold-start	53.6	77.0	75.3	38.7	24533.7	723.1	11.0
Re-Omni	**54.6**	**78.7**	**76.3**	**42.9**	**24030.1**	**101.7**	**9.5**

*For latest video, the input entry is the peak context retained simultaneously, averaged over questions. Each question involves an average of 3.8 viewing rounds; this entry is not cumulative processing and is therefore not comparable to a cumulative input-token count.

5.3 Analysis of Re-Ask

Table 3 compares three separately trained settings that vary the repeated question $\mathbf{Q}'$ and revise reward r_{rev}. We evaluate initial and refined answers on each full benchmark test set. For a complementary attention analysis, we randomly sample 50 test examples from VideoMME and sum the attention assigned by the re-answer token across layers, heads, and the token positions of the original question $\mathbf{Q}$, initial answer $\mathbf{A}^{(1)}$, and repeated question $\mathbf{Q}'$; the table reports the mean mass for each anchor across the samples.

Table 3: Initial- and refined-answer accuracy and re-answer attention mass under three separately trained RL settings. Accuracy uses each full benchmark test set; attention mass averages over 50 randomly sampled examples.

Setting	Pass	Accuracy				Attention Mass		
		VideoHolmes	DailyOmni	VideoMME	LVOmniBench	$\mathbf{a}_{\mathbf{Q}}$	$\mathbf{a}_{\mathbf{A}^{(1)}}$	$\mathbf{a}_{\mathbf{Q}'}$
w/o $\mathbf{Q}'$, w/ r_{rev}	$\mathbf{A}^{(1)}$	52.6	77.3	71.9	38.5	0.09	0.06	-
	$\mathbf{A}^{(2)}$	52.7	77.3	71.9	38.5			
w/ $\mathbf{Q}'$, w/o r_{rev}	$\mathbf{A}^{(1)}$	53.0	78.0	73.6	42.6	0.10	0.08	0.04
	$\mathbf{A}^{(2)}$	53.0	77.9	73.6	42.6			
w/ $\mathbf{Q}'$, w/ r_{rev}	$\mathbf{A}^{(1)}$	53.1	76.3	73.6	40.8	0.03	0.02	0.24
	$\mathbf{A}^{(2)}$	**54.6**	**78.7**	**76.3**	**42.9**			

The two single-component settings show almost no answer-stage gain: with r_{rev} but without $\mathbf{Q}'$, and with $\mathbf{Q}'$ but without r_{rev}, the change from $\mathbf{A}^{(1)}$ to $\mathbf{A}^{(2)}$ stays within 0.1 percentage points on all four benchmarks. With both components enabled, the refined answer improves by 1.5, 2.4, 2.7, and 2.1 points on VideoHolmes, DailyOmni, VideoMME, and LVOmniBench, respectively. On DailyOmni and LVOmniBench, the full setting starts with a less accurate initial answer than the setting without r_{rev} but finishes with a more accurate final answer; its advantage therefore comes from the second-pass revision rather than a stronger first pass. The repeated question places a query after the new evidence, while the revise reward discourages retaining an incorrect initial answer. Their joint use is associated with the answer-stage gains; neither component alone produces comparable improvement in these runs. In the 50-example attention probe, mean mass on $\mathbf{Q}'$ increases from 0.04 without r_{rev} to 0.24 in the full setting, alongside lower mass on the original question and initial answer. This attention pattern is consistent with greater use of the repeated question, but does not by

itself explain the accuracy gain. Appendix E further evaluates replacing the repeated question with a generic prompt in the final Re-Omni model. This suggests that Re-Omni learns to use the new video evidence beyond relying on exact question repetition, while re-asking the original question provides additional question-specific guidance.

The VideoMME answer transitions show how the full setting improves: 7.1% of examples change from wrong to right, 4.4% from right to wrong, 4.1% from wrong to another wrong option, and 84.4% keep the same answer. Successful corrections exceed regressions by 2.7 percentage points, matching the accuracy increase from 73.6% to 76.3% between passes. Thus, the gain reflects a positive balance of corrections over newly introduced errors, rather than simply more frequent answer changes.

5.4 Analysis of Re-Watch and Re-Answer

Table 4 compares separately trained variants that change the second-pass evidence or answer structure. *Re-Answer only* keeps the full model's RL settings but repeats answering without new video frames. *Uniform Re-Watch* uses the same second-pass visual-token budget to sample across the full video instead of the selected interval. $\mathbf{A}^{(2)}$ *only* retains targeted re-watching but omits the initial answer and trains with r_{fmt1}, r_{fmt2}, and r_{acc2}. Single-pass QA-SFT provides a supervised reference.

Table 4: Final-answer accuracy for separately trained re-watch and re-answer variants. Re-Answer only adds no new video frames; Uniform Re-Watch samples the full video under the same second-pass visual-token budget; $\mathbf{A}^{(2)}$ only removes the initial answer and its associated accuracy and revise rewards.

Variant	VideoHolmes	DailyOmni	VideoMME	LVOmniBench
QA-SFT	53.4	77.9	72.9	40.6
Re-Answer only	53.1	77.3	74.4	39.7
Uniform Re-Watch	52.6	78.4	73.9	41.0
$\mathbf{A}^{(2)}$ only	53.9	77.4	73.1	39.6
Re-Omni	**54.6**	**78.7**	**76.3**	**42.9**

Re-Omni improves over Re-Answer only by 1.5, 1.4, 1.9, and 3.2 percentage points on VideoHolmes, DailyOmni, VideoMME, and LVOmniBench, respectively. Re-Answer-only also trails QA-SFT on three benchmarks, so a second opportunity to answer without new frames does not reliably improve accuracy. Adding the selected video changes this outcome, with the largest gain on LVOmniBench. Using the selected interval instead of uniform sampling under the same visual-token budget raises accuracy by 2.0, 0.3, 2.4, and 1.9 points, respectively. The margins on VideoHolmes and both long-video benchmarks indicate that where the second-pass budget is spent matters, beyond the presence of additional high-fidelity frames alone. DailyOmni is a weaker case for targeted selection: its 0.3-point margin means uniform re-watching nearly matches the selected clip on that benchmark. Finally, Re-Omni exceeds $\mathbf{A}^{(2)}$ only by 0.7, 1.3, 3.2, and 3.3 points; the larger gaps on the long-video benchmarks show that removing the initial answer and its associated rewards is particularly costly there. Together, these controls support the combined roles of new evidence, targeted selection, and the full answer-first design in final-answer accuracy. Appendix G provides qualitative examples of successful refinement.

6 Conclusion

Re-Omni enables two-pass video understanding through direct RL from an instruction-tuned baseline, without CoT cold-start, multi-turn re-watch, or explicit second-pass reasoning. Answer-level and format rewards train interval selection: **re-watch** acquires higher-spatial-resolution evidence, **re-ask** repeats the question afterward, and **re-answer** refines the initial direct answer. Re-Omni improves over same-data QA-SFT on all six benchmarks and outperforms both multi-turn variants with two viewing rounds and 101.7 average output tokens. Matched-budget targeted-versus-uniform and answer-stage ablations support the combined workflow. The extra view costs more than single-pass QA-SFT, but avoids the long output traces and higher latency of the evaluated multi-turn recipes while improving accuracy in the reported comparisons.

AI USE DISCLOSURE

In this work, we used generative AI tools to assist with portions of the software implementation and to polish the manuscript for clarity, grammar, and readability. The authors reviewed and revised all AI-assisted outputs and take full responsibility for the final content of this work, including its code, text, claims, and other artifacts.

REFERENCES

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common Voice: A Massively-Multilingual Speech Corpus. In *Proc. LREC*, 2020.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025.

Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. CG-Bench: Clue-grounded Question Answering Benchmark for Long Video Understanding. In *Proc. ICLR*, Singapore, 2025.

Xinlong Chen, Yue Ding, Weihong Lin, Jingyun Hua, Linli Yao, Yang Shi, Bozhou Li, Yuanxing Zhang, Qiang Liu, Pengfei Wan, et al. AVoCaDO: An Audiovisual Video Captioner Driven by Temporal Orchestration. In *Proc. ICLR*, Rio de Janeiro, 2026.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-Holmes: Can MLLM Think Like Holmes for Complex Video Reasoning? *arXiv preprint arXiv:2505.21374*, 2025.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*, 2024.

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.

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.

Yang Ding, Yizhen Zhang, Xin Lai, Ruihang Chu, and Yujiu Yang. VideoZoomer: Reinforcement-Learned Temporal Focusing for Long Video Reasoning. In *Proc. ICLR*, Rio de Janeiro, 2026.

Junjie Fei, Jun Chen, Zechun Liu, Yunyang Xiong, Chong Zhou, Wei Wen, Junlin Han, Mingchen Zhuge, Saksham Suri, Qi Qian, et al. Small Vision-Language Models are Smart Compressors for Long Video Understanding. *arXiv preprint arXiv:2604.08120*, 2026.

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. In *Proc. NeurIPS*, Mexico, 2025.

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 *Proc. CVPR*, Nashville TN, 2025.

Shenghao Fu, Qize Yang, Yuan-Ming Li, Xihan Wei, Xiaohua Xie, and Wei-Shi Zheng. LOVE-R1: Advancing Long Video Understanding with an Adaptive Zoom-in Mechanism via Multi-Step Reasoning. In *Proc. ICLR*, Rio de Janeiro, 2026.

Hong Gao, Yiming Bao, Xuezhen Tu, Bin Zhong, Linan Yue, and Min-Ling Zhang. APVR: Hour-Level Long Video Understanding with Adaptive Pivot Visual Information Retrieval. In *Proc. AAAI*, Singapore, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *Nature*, 645: 633–638, 2025.

Pin-Lun Hsu, Yun Dai, Vignesh Kothapalli, Qingquan Song, Shao Tang, Siyu Zhu, Steven Shimizu, Shivam Sahni, Haowen Ning, Yanning Chen, and Zhipeng Wang. Liger-Kernel: Efficient Triton Kernels for LLM Training. In *Proc. ICML Workshop on CODEML*, Vancouver, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *Proc. ICLR*, 2022.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating Captions for Audios in the Wild. In *Proc. NAACL-HLT*, 2019.

Minsoo Kim, Kyuhong Shim, Jungwook Choi, and Simyung Chang. InfiniPot-V: Memory-Constrained KV Cache Compression for Streaming Video Understanding. *arXiv:2506.15745*, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proc. SOSP*, Koblenz, 2023.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Jiafu Tang, Zhenghao Song, Dingling Zhang, et al. OmniVideoBench: Towards Audio-Visual Understanding Evaluation for Omni MLLMs. In *Proc. ICLR*, Rio de Janeiro, 2026.

Yixuan Li, Changli Tang, Jimin Zhuang, Yudong Yang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. Improving LLM Video Understanding with 16 Frames Per Second. In *Proc. ICML*, Vancouver, 2025.

Xuyang Liu, Yiyu Wang, Junpeng Ma, and Linfeng Zhang. Video Compression Commander: Plug-and-Play Inference Acceleration for Video Large Language Models. *arXiv preprint arXiv:2505.14454*, 2025a.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. NVILA: Efficient Frontier Visual Language Models. In *Proc. CVPR*, Nashville, 2025b.

Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, et al. Omni-Captioner: Data Pipeline, Models, and Benchmark for Omni Detailed Perception. In *Proc. ICLR*, Rio de Janeiro, 2026.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. WavCaps: A ChatGOT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research. *IEEE Transactions on Audio, Speech and Language Processing*, 32: 3339–3354, 2024.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR Corpus Based on Public Domain Audio Books. In *Proc. ICASSP*, 2015.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming Long Video Understanding with Large Language Models. In *Proc. NeurIPS*, 2024.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling Video LLMs with Active Real-Time Interaction via Disentangled Perception, Decision, and Reaction. In *Proc. CVPR*, 2025.

Minghao Qin, Xiangrui Liu, Zhengyang Liang, Yan Shu, Huaying Yuan, Juenjie Zhou, Shitao Xiao, Bo Zhao, and Zheng Liu. Video-XL-2: Towards Very Long-Video Understanding Through Task-Aware KV Sparsification. *arXiv preprint arXiv:2506.19225*, 2025.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-scale Weak Supervision. In *Proc. ICML*, 2023.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters. In *Proc. KDD*, 2020.

Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. CinePile: A Long Video Question Answering Dataset and Benchmark. *arXiv:2405.08813*, 2024.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, Yuxuan Wang, and Chao Zhang. video-SALMONN: Speech-Enhanced Audio-Visual Large Language Models. In *Proc. ICML*, 2024.

Guangzhi Sun, Yixuan Li, Xiaodong Wu, Yudong Yang, Wei Li, Zejun Ma, and Chao Zhang. video-SALMONN S: Memory-Enhanced Streaming Audio-Visual LLM. *arXiv preprint arXiv:2510.11129*, 2025.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. SALMONN: Towards Generic Hearing Abilities for Large Language Models. In *Proc. ICLR*, Vienna, 2024.

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.

Changli Tang, Tianyi Wang, Fengyun Rao, Jing LYU, and Chao Zhang. D-ORCA: Dialogue-Centric Optimization for Robust Audio-Visual Captioning. *arXiv preprint arXiv:2602.07960*, 2026.

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.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Xiao Wang, Qingyi Si, Jianlong Wu, Shiyu Zhu, Li Cao, and Liqiang Nie. AdaReTaKe: Adaptive Redundancy Reduction to Perceive Longer for Video-language Understanding. In *Proc. ACL (Findings)*, Vienna, 2025b.

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

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

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 *Proc. EMNLP*, Suzhou, 2025.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Chengwei Qin, Shijian Lu, Xingxuan Li, and Lidong Bing. LongVT: Incentivizing "Thinking with Long Videos" via Native Tool Calling. In *Proc. CVPR*, Colorado, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *Proc. ICLR*, Kigali, 2023.

Jeong Hun Yeo, Sangyun Chung, Sungjune Park, Dae Hoe Kim, Jinyoung Moon, and Yong Man Ro. GCAgent: Long-video understanding via schematic and narrative episodic memory. *arXiv preprint arXiv:2511.12027*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, et al. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. In *Proc. NeurIPS*, Mexico, 2025.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. *arXiv preprint arXiv:2501.13106*, 2025. URL https://arxiv.org/abs/2501.13106.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-VStream: Memory-Based Real-Time Understanding for Long Video Streams. *arXiv preprint arXiv:2406.08085*, 2024a.

Yiming Zhang, Zhuokai Zhao, Zhaorun Chen, Zenghui Ding, Xianjun Yang, and Yining Sun. Beyond Training: Dynamic Token Merging for Zero-Shot Video Understanding. *arXiv preprint arXiv:2411.14401*, 2024b.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video Instruction Tuning with Synthetic Data. *arXiv preprint arXiv:2410.02713*, 2024c.

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.

Jialong Zuo, Yongtai Deng, Lingdong Kong, Jingkang Yang, Rui Jin, Yiwei Zhang, Nong Sang, Liang Pan, Ziwei Liu, and Changxin Gao. VideoLucy: Deep Memory Backtracking for Long Video Understanding. In *Proc. NeurIPS*, Mexico, 2025.

A LIMITATIONS

Our evaluation covers one 8B model family and multiple-choice benchmarks; the results do not establish the same behavior for other model scales, architectures, or open-ended generation. The primary QA-SFT/Re-Omni comparison uses only three paired seeds, making its Student-t intervals imprecise, while the other configurations and ablations remain single-run results. The attention analysis is correlational, and its 50-example subset cannot identify the mechanism behind the answer-stage gains. Finally, temporal intervals are optimized indirectly through answer and format rewards; downstream answer gains and the targeted-versus-uniform ablation show that the selected clips are useful to the workflow, but do not establish localization quality under a dedicated temporal metric.

B MORE IMPLEMENTATION DETAILS

B.1 SCOPE OF RECIPE COMPARISONS

The comparisons evaluate complete training and interaction recipes. Different interaction protocols require different prompts and output structures: Direct RL uses re-watch alone, without repeating the question or producing an initial answer for refinement, whereas Re-Omni adds re-ask and answer-first re-answer. Requiring identical interaction fields would remove the interventions under study. Table 4 provides a matched RL control: Re-Answer only uses the same initialization, LoRA scaling, and training settings as full Re-Omni, with re-watching disabled. It repeats answering without new visual evidence. The full model's gains over this control support the contribution of re-watching within the evaluated workflow; the control does not measure temporal localization accuracy.

B.2 SYSTEM PROMPT

The reinforcement learning stage of Re-Omni relies on a carefully crafted system prompt that specifies the protocol in a machine-parseable schema. Rather than distilling chain-of-thought trajectories from a teacher model and cold-starting the policy via SFT, we provide only a minimal format prior through the system message below, and let the model explore full trajectories under rule-based rewards. The exact system message used in both training and inference is shown in Fig. 2.

The training-time validity check operates on the parsed interval rather than on the prompt text. It uses the first-pass sampling rate and the 128-frame second-pass budget to reject intervals that would reduce temporal sampling density; for videos already sampled at 10 FPS in the first pass, equal temporal density is accepted. Intervals covering the full video are rejected.

The prompt is structured around three loci of model output. In the first pass, the *initial answer* is emitted *before* any `<thinking>` or `<tool_call>` tokens. Consequently, nothing precedes $\mathbf{A}^{(1)}$ except $(\mathbf{V}^{(1)}, \mathbf{Q})$, preserving the instruction-tuned model's question-to-answer prefix interface without inserting a reasoning scaffold. The subsequent `<thinking>` block provides a free-form slot for explaining why an interval is relevant, and the terminal `<tool_call>locate [start, end]</tool_call>` yields a parseable interval that drives second-pass sampling. In the second pass, the prompt asks for a clean re-answer that either repeats $\mathbf{A}^{(1)}$ or corrects it, paired with the r_{fmt2} and r_{rev} reward terms defined in the main text.

B.3 LORA SCALING BETWEEN SFT AND RL

During RL, only rank-128 LoRA adapters on all query, key, and value projection layers of the LLM are trainable. The base-model weights and all other modules remain frozen.

Before end-to-end RL, we change the LoRA scaling factor following the practice described in Tang et al. (2024). During audio-visual caption SFT, the adapter uses $\alpha_{\text{SFT}} = 256$. At the start of RL, we set $\alpha_{\text{RL}} = 32$ while keeping the learned low-rank matrices A and B unchanged. Since the effective LoRA contribution is

$$\Delta W = \frac{\alpha}{r} BA, \tag{7}$$

reducing α from 256 to 32 attenuates the adapter contribution by a factor of 8 before RL begins, without changing the direction represented by BA.

```
You are a helpful assistant.
###
In the first round, you will be given a full video along with the
question.
FIRST: Output your initial answer.
THEN: Think carefully to find a timestamp that is related to the
question.  Output the reasoning progress as an internal monologue
enclosed within <thinking>...</thinking>.  Your reasoning progress
should follow the following logic:
The video generally describes ...(overall description of the video).
The question asks about ...(specific detail requested).  To answer
this question, I should look for segments showing ...(what needs to
be found).  In the video, the time segment ...  is related to this
question because ...(reason).  Therefore, I should locate this part as
it contains the answer...
AT LAST: Output the located timestamp in second with the format
<tool_call>locate [start, end]</tool_call>
Output format:
...<thinking>...</thinking><tool_call>locate [start, end]</tool_call>

###
In the second round, you will be additionally given a segmented video
clip with higher fidelity.  You should directly output the final
answer based on both the full video and the located clip.  If you
believe the previous answer was correct, repeat the initial answer;
otherwise, correct it.
```

Figure 2: The exact system prompt used for end-to-end RL in Re-Omni. The first-round specification defines a three-part output (*initial answer* → `<thinking>` reasoning → `<tool_call>locate` call), while the second-round specification constrains the re-answer to be a clean final answer that either repeats or corrects the initial one.

This asymmetric $\alpha_{\text{SFT}} > \alpha_{\text{RL}}$ schedule retains the learned adapter direction while reducing its magnitude at the RL entry point. It therefore initializes RL from a rescaled version of the caption-SFT adapter rather than from either the full SFT scaling or an unadapted backbone.

To assess the effect of scaling alone, we evaluate QA-SFT with its LoRA scaling changed from $\alpha = 256$ to $\alpha = 32$ on the four benchmarks used for the workflow ablations. Table 5 reports this control alongside the original QA-SFT and the evaluated Re-Omni checkpoint. Rescaled QA-SFT obtains 52.9, 77.8, 72.9, and 40.7, differing from the original QA-SFT by -0.5, -0.1, 0.0, and $+0.1$ percentage points, respectively. The corresponding Re-Omni scores are higher by 1.7, 0.9, 3.4, and 2.2 points. Thus, the scaling change alone does not reproduce the Re-Omni gains in these single-run comparisons.

Table 5: Effect of LoRA scaling on QA-SFT. Accuracy is reported on the four workflow-ablation benchmarks for single evaluated runs. The QA-SFT + scaling row applies the RL-stage scaling factor without RL training.

Setting	VideoHolmes	DailyOmni	VideoMME	LVOmniBench
QA-SFT ($\alpha = 256$)	53.4	77.9	72.9	40.6
QA-SFT + LoRA scaling ($\alpha = 32$)	52.9	77.8	72.9	40.7
Re-Omni	54.6	78.7	76.3	42.9

B.4 Engineering Implementation Details

The end-to-end RL stage uses the following engineering configuration.

Rollout generation with vLLM. Each RL optimization step uses a group of G rollouts per question. We use vLLM (Kwon et al., 2023) as the rollout engine for paged-attention-based KV-cache management and continuous batching. The generated trajectories have variable lengths under the two-pass protocol, so completed trajectories can leave the batch while longer trajectories continue decoding.

Training with Liger-Kernel. For the forward-backward pass, we use Liger-Kernel (Hsu et al., 2025) to replace several memory- and compute-heavy operators in the Qwen3-VL backbone with fused Triton kernels, including cross-entropy, RMSNorm, RoPE, and SwiGLU. Token log-probabilities for the DAPO loss are computed from the full-vocabulary softmax at temperature 0.95; rollout sampling uses temperature 0.95, top-k 20, and top-p 0.95. We combine these kernels with DeepSpeed ZeRO-1 (Rasley et al., 2020) for optimizer-state sharding when training trajectories containing on the order of 20k visual tokens.

Colocate mode with bidirectional offloading. We run the rollout engine and trainer in colocate mode on the same GPUs, with bidirectional CPU offloading: during the forward-backward pass, the vLLM engine (including its weights and KV cache) is offloaded to CPU to free GPU memory for training; during generation, the trainer's model parameters, gradients, and optimizer states are offloaded to CPU, freeing GPU memory for vLLM's paged KV cache. This alternating schedule, paired with ZeRO-1, allows a single H800 to accommodate both roles without the parameter duplication of a separate-GPU actor–learner split.

Per-GPU rollout accumulation. Every rollout requires decoding and resampling raw video frames. To avoid decoding the same sample independently on every GPU, we reshape the data-parallel layout: instead of broadcasting one sample to all GPUs and running G rollouts in parallel, each GPU processes one question and runs `accum_step` sequential rollouts, after which the rollouts are all-gathered across ranks. The resulting effective batch size is the world size (the number of GPUs), and the effective group size is `accum_step`, matching the group-normalized advantage in Eqn. (5). This layout avoids replicated decoding of identical videos across ranks and distributes the decoding load across the cluster.

B.5 Resources

Audio alignment and audio-visual captioning SFT use 32 H800s for about 10 hours. The RL stage uses 96 H800s for about 48 hours; excluding video-decoding time, its training operations take about 30 hours. Asynchronous video decoding is a potential direction for reducing the remaining decoding overhead.

C Reward Curves

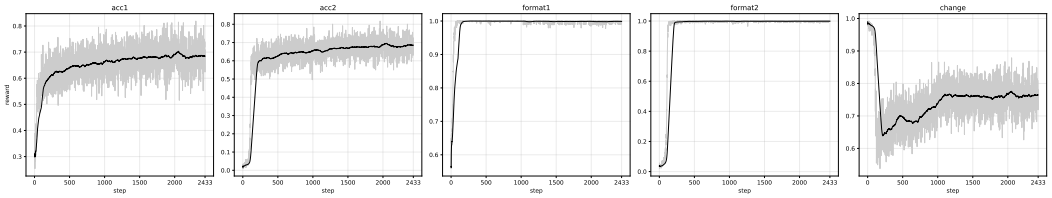

Figure 3: Logged reward curves during training, with optimization step on the horizontal axis and reward value on the vertical axis. Each plotted value is the arithmetic mean of the corresponding per-trajectory rewards at that logged step; no additional curve smoothing is applied. Panels `acc1` and `acc2` correspond to the initial- and refined-answer accuracy rewards r_{acc1} and r_{acc2}; `format1` and `format2` correspond to r_{fmt1} and r_{fmt2}; and `change` corresponds to the revise reward r_{rev}. For trajectories reaching the second pass, $r_{\text{rev}} = 1$ except when an incorrect initial answer is repeated, so `change` is neither a correction rate nor a direct measure of successful refinement.

Figure 3 shows that both format rewards rise rapidly and remain near one, while the two accuracy rewards increase over training. The `change` reward drops early and then partially recovers. Answer-stage accuracy before and after the second pass is reported in Table 3.

D More Ablations on RL Settings

We further conduct additional experiments to examine the sensitivity of our recipe to key RL hyperparameters and loss formulations, with results summarized in Table 6.

Table 6: Ablation on RL settings. We compare group sizes and GRPO with the DAPO objective in Eqn. (6). The default configuration uses group size 8.

Setting	VideoHolmes	DailyOmni	VideoMME	LVOmniBench
group size=4	**55.0**	77.8	75.4	40.7
GRPO loss	53.9	77.4	73.1	40.2
Ours (G=8, DAPO)	54.6	**78.7**	**76.3**	**42.9**

E Re-Ask with a Generic Prompt

We evaluate the final Re-Omni model with the repeated question $\mathbf{Q}'$ replaced at inference time by the generic prompt "What is the answer for the question?" Table 7 compares the initial answer, the standard refined answer, and the refined answer with this replacement.

Table 7: Accuracy (%) of the final Re-Omni model with the original question or a generic prompt in the second pass.

Answer	VideoHolmes	DailyOmni	VideoMME	LVOmniBench
$\mathbf{A}^{(1)}$	53.1	76.3	73.6	40.8
$\mathbf{A}^{(2)}$ (repeated question)	**54.6**	**78.7**	**76.3**	**42.9**
$\mathbf{A}^{(2)}$ (generic prompt)	54.4	78.1	76.0	41.9

The generic prompt still improves over $\mathbf{A}^{(1)}$ on all four benchmarks, suggesting that Re-Omni training improves the model's ability to use the second-pass video evidence even without repeating the original question. Repeating the same question nevertheless yields consistently higher accuracy.

F Seed-Wise Primary Comparison

Re-Omni improves on QA-SFT in all 18 matched seed–benchmark comparisons. The paired gains are 1.1 ± 0.2, 0.5 ± 0.3, 3.4 ± 0.0, 2.9 ± 0.4, 2.9 ± 0.6, and 2.4 ± 0.2 percentage points in (mean $\pm$ sample standard deviation across the three paired seeds). The individual runs are reported below. Two-sided 95% Student-t intervals for the mean paired gains ($n = 3$, $df = 2$) are $[0.7, 1.5]$, $[-0.1, 1.2]$, $[3.4, 3.4]$, $[1.8, 4.0]$, $[1.3, 4.5]$, and $[2.0, 2.8]$ percentage points, respectively. Five intervals exclude zero, whereas the DailyOmni interval does not; with only three paired runs, these estimates remain imprecise.

Table 8: Individual runs underlying the mean and sample standard deviation in Table 1. QA-SFT and Re-Omni use matched seeds.

Model	**Seed**	**VideoHolmes**	**DailyOmni**	**AVUT**	**OmniVideoBench**	**VideoMME**	**LVOmniBench**
QA-SFT	2026	53.4	77.9	74.1	40.8	72.9	40.6
	2025	53.3	77.9	74.2	40.8	72.7	40.5
	2024	53.6	77.9	74.1	40.8	72.9	40.4
Re-Omni	2026	54.6	78.7	77.5	43.9	76.3	42.9
	2025	54.5	78.4	77.6	43.2	75.9	42.8
	2024	54.5	78.2	77.5	44.0	75.1	43.0

G SELECTED QUALITATIVE REFINEMENT CASES

Tables 9–11 present three selected successful refinements that illustrate the observed two-pass behavior.

Table 9: Case 1 illustrates re-watching a relevant segment. The final answer $\mathbf{A}^{(2)} = \mathrm{A}$ matches the ground truth.

Video Source	**Question**	$\mathbf{A}^{(1)}$	$\mathbf{R}^{(1)}$	$\mathbf{T}$	$\mathbf{A}^{(2)}$
VideoMME FsLaTZmP6Uw	Which sentence describes this rally according to this video? A. Two athletes make a series of consecutive successful hits. B. The athlete in red can not easily make a good return. C. The athlete in black often makes edge balls. D. None of the above is correct.	B	The video generally describes a table tennis rally between two athletes, one in a red shirt and one in a black shirt, with the score displayed as CHN 3-2 JPN. The question asks about a specific detail in this rally. To answer this, I should look for segments showing the difficulty of returns by the athlete in red. In the video, the athlete in red is seen making multiple attempts to return the ball, often stretching or diving, indicating difficulty. The time segment 14.2-33.8 is related to this question because it shows the entire rally where the athlete in red struggles to return the ball effectively, while the athlete in black consistently makes powerful returns. Therefore, I should locate this part as it may be helpful.	[14.2, 33.8]	A

Table 10: Case 2 illustrates inspecting localized evidence at higher fidelity. The final answer $\mathbf{A}^{(2)} = \mathrm{A}$ matches the ground truth.

Video Source	**Question**	$\mathbf{A}^{(1)}$	$\mathbf{R}^{(1)}$	$\mathbf{T}$	$\mathbf{A}^{(2)}$
LVOmniBench video_72	During the first segment of background music, by looking in the mirror, which hand can we infer the vlogger is using to hold the camera? A. Right hand. B. Left hand. C. It cannot be determined. D. Both hands.	B	The video generally describes a vlogger exploring a new home and showcasing its features. The question asks about a specific detail regarding the vlogger's hand position while looking in the mirror. To answer this question, I should look for segments showing the vlogger looking in the mirror and holding the camera. In the video, at 65s, the vlogger is seen looking in the mirror while holding the camera, and it is evident that she is using her left hand to hold the camera.	[61.3, 68.8]	A

Table 11: Case 3 illustrates re-watching a localized clip to refine the answer. The final answer $\mathbf{A}^{(2)} = \mathrm{E}$ matches the ground truth.

Video Source	**Question**	$\mathbf{A}^{(1)}$	$\mathbf{R}^{(1)}$	$\mathbf{T}$	$\mathbf{A}^{(2)}$
VideoHolmes d4Gx0UTdFiU	What is the girl's intention of adding water to nail polish in the kitchen? A. Dilute the overly thick texture. B. Test the mixing effect of liquids. C. Extend the usage time of nail polish. D. Imitate the steps of beauty tutorial. E. Pre-meditated to cover up the color of bloodstains. F. A spur-of-the-moment prank	A	The video generally describes a girl walking into a room, removing her shoes, and then sitting on the floor to apply nail polish to her feet. The question asks about her intention of adding water to the nail polish in the kitchen. To answer this question, I should look for segments showing her mixing water with nail polish. In the video, the time segment 1.5s-97.7s is related to this question because it shows the girl walking into the room, removing her shoes, and then sitting on the floor to apply nail polish to her feet. Therefore, I should locate this part as it may be helpful.	[1.5, 97.7]	E