# RETHINKING AUDIO-VISUAL TEMPORAL ALIGNMENT WITH SELF-SUPERVISED REINFORCEMENT LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Audio-visual video understanding has become a foundational task in multimodal learning, supporting applications from embodied perception to audio-conditioned video generation, and requires models to jointly reason over temporally aligned audio and visual streams. At the core of this task lies temporal alignment, the binding of audio events to their corresponding visual moments at sub-second granularity. Current omni-modal models often fail at it, a problem commonly referred to as *temporal misalignment*, suggesting that they have only achieved coarse-grained audio-visual alignment instead of fine-grained temporal correspondence. To address this, we propose a **self-supervised reinforcement learning** framework that supervises temporal alignment without external annotations or auxiliary models. We insert paired audio and visual events at controlled onsets into real videos, yielding **AVABench**, an alignment-focused benchmark whose ground truth is defined entirely by the insertion process. The model is then trained with reinforcement learning directly on these construction-derived signals, avoiding the annotation cost of prior supervised pipelines and directly targeting the temporal binding that supervised captions leave unaddressed. Experiments show that existing models achieve only 39.15% accuracy on AVABench, while our 3B **Omni-Aligner** reaches 74.77%, surpassing a 7B baseline by over 20 points, and the learned capability transfers consistently to multiple out-of-domain audio-visual benchmarks.

## 1 INTRODUCTION

Audio-visual video understanding has become a foundational task in multimodal learning, supporting downstream applications from embodied perception to audio-conditioned video generation (Li et al., 2025b; Xu et al., 2025b; Tang et al., 2025; Ye et al., 2025b; Geng et al., 2025a). Beyond determining *what* occurs in each modality, the task demands precise temporal alignment, where models must bind audio events to their corresponding visual moments at sub-second granularity, so that a sound is attributed to the specific action that produced it rather than to another event that merely co-occurs nearby. Current omni-modal models frequently fail at this, a problem commonly referred to as *temporal misalignment* that reflects reliance on coarse audio-visual co-occurrence rather than precise temporal correspondence. The severity of this failure is empirically pronounced, with alignment sub-categories consistently ranking lowest across established audio-visual benchmarks, and even the strongest open-source omni-model in our study attaining only 42.5% on direct cross-modal binding, barely above the 25% random-chance baseline. Notably, unlike classical clip-level synchronization tasks that reduce to binary judgments over short segments, this misalignment arises during fine-grained, open-ended reasoning within long-form videos, where the model must precisely localize *when* a coupled event occurs and verify *whether* the two streams truly correspond at that moment, rather than merely co-occur somewhere in the clip.

Most existing approaches construct audio-visual instruction data by segmenting videos, extracting uni-modal captions, and synthesizing them into QA pairs (Ye et al., 2024; Liu et al., 2025b; Geng et al., 2025b; Ye et al., 2025a; Kurpath et al., 2025), then perform end-to-end supervised fine-tuning that demands costly high-quality annotations. Subsequent efforts mitigate narrative incoherence through entity-anchored scripting (Cai et al., 2026) or full-video revision (Zhou et al., 2025), and recent works apply reinforcement learning via self-supervised temporal reordering (Jia et al., 2026) or contrastive fusion rewards (Chen et al., 2026b), yet none provide an explicit temporal alignment objective. Moreover, their reliance on token-level imitation over synthesized references limits ex-



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.

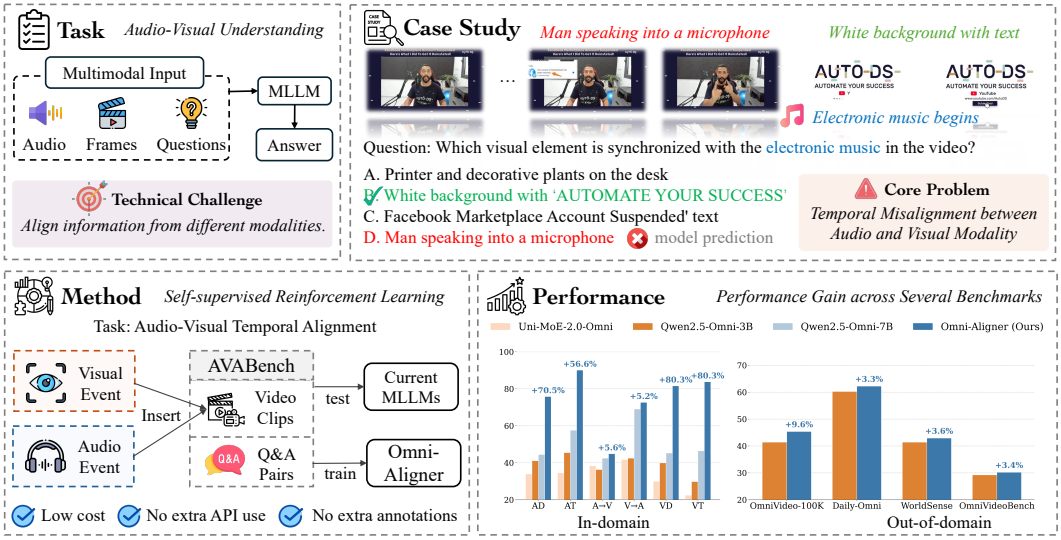


Figure 1: **Overview of Our Proposed Method.** We propose a self-supervised reinforcement learning framework that addresses audio-visual temporal misalignment in multimodal LLMs, training the low-cost AVABench dataset to yield Omni-Aligner. Omni-Aligner achieves consistent gains over existing baselines across in-domain and out-of-domain benchmarks.

ploration and scales poorly with real audio-visual annotation cost, leaving fine-grained temporal binding largely unaddressed.

We propose a self-supervised reinforcement learning framework that directly targets temporal misalignment without extra annotations or external models. Our key idea follows a capability-not-distribution principle: we insert paired visual and audio events at a shared onset into real videos, using primitives that are trivially recognizable so the reward gradient is forced to concentrate on *when* the two streams should be bound rather than *what* they contain. This construction yields **AVABench**, an alignment-focused benchmark with 31,188 training and 3,024 test samples whose ground-truth coupling derives entirely from insertion metadata under strict video-level non-overlap. On top of it, we adopt a GRPO objective that rewards structured reasoning trajectories with composite signals over timestamp accuracy, format compliance, and answer correctness, replacing token-level imitation with reward-driven exploration. Existing models achieve only 39.15% accuracy on AVABench, while our 3B **Omni-Aligner** reaches 74.77%, surpassing a 7B baseline by over 20 points. Notably, pure GRPO outperforms both SFT and SFT followed by GRPO, and gains transfer consistently to multiple out-of-domain audio-visual benchmarks, confirming that once temporal binding is explicitly targeted as a capability, it generalizes well beyond the training vocabulary.

Our contributions are summarized as follows:

- **Problem and benchmark.** We identify *time misalignment* as a prevalent failure mode in audio-visual understanding, empirically confirmed across multiple benchmarks, and introduce **AVABench**, an alignment-focused benchmark whose ground truth is obtained by construction, requiring no extra annotations or external models.

- **Self-supervised RL framework.** We propose a GRPO-based reinforcement learning framework that trains the model to explore temporal alignment through self-generated reasoning trajectories, avoiding the overfitting and limited generalization of token-level supervised fine-tuning.

- **Effectiveness and generalization.** **Omni-Aligner** improves accuracy from 39.15% to 74.77% on AVABench and consistently outperforms its backbone on multiple out-of-domain audio-visual benchmarks, confirming that mitigating temporal misalignment benefits broader audio-visual comprehension.

## 2 Related Work

**Multimodal Large Language Models.** Prior Multimodal Large Language Models (MLLMs) have achieved notable progress in visual understanding across image (Wu et al., 2025b; Liu et al., 2025a; Fu et al., 2026a) and video (Wu et al., 2025a; Sun et al., 2026; Wu et al., 2026b; Ge et al., 2026; Fu et al., 2026b) domains (Li et al., 2024; Bai et al., 2025; Marafioti et al., 2025; Wang et al., 2025). To tackle complex tasks, research has progressively shifted from passive perception toward eliciting multimodal reasoning. Current approaches primarily rely on explicit chain-of-thought (CoT) (Wei et al., 2022; Yang et al., 2025) trajectories or reinforcement learning with verifiable rewards (RLVR) (Feng et al., 2026; Ge et al., 2025; Wu et al., 2026a) to incentivize step-by-step deduction and self-verification. However, seamlessly extending these reasoning paradigms to dynamically coordinate continuous sensory streams remains an open challenge.

**Audio-Visual Understanding.** Extensive audio-visual benchmarks (Alamri et al., 2019; Yun et al., 2021; Li et al., 2022; Yang et al., 2022; Hong et al., 2025; Li et al., 2025a; Tao et al., 2026; Li et al., 2025d) have been developed to evaluate cross-modal reasoning. Methodologically, segment-and-caption pipelines (Geng et al., 2025b; Chao et al., 2025; Ye et al., 2025a) often suffer from narrative incoherence, which OmniVideo-100K (Cai et al., 2026) and Daily-Omni (Zhou et al., 2025) alleviate via clue-guided QA generation and full-video context revision. Beyond data curation, recent works employ reinforcement learning to enhance audio-visual reasoning, including self-supervised temporal reordering in OmniJigsaw (Jia et al., 2026) and modality-attentive fusion rewards in OmniVideo-R1 (Chen et al., 2026b). To overcome the information loss in discrete textual chains-of-thought, LatentOmni (Dai et al., 2026) interleaves text with synchronized continuous audio-visual latent states. For long-horizon understanding, OmniRAG-Agent (Zhu et al., 2026) leverages multi-turn retrieval-augmented tool calling, while OmniReasoner (Chen et al., 2026a) enables models to dynamically zoom into high-fidelity audio-video segments from a coarse global preview via TimeAnchor.

**Self-supervised learning for Video Understanding** Self-supervised learning (Noroozi & Favaro, 2016; Xu et al., 2019; Wang et al., 2020; Schiappa et al., 2023) for video learns spatio-temporal representations from unlabeled data via intrinsic pretext tasks. While early pretext tasks like video jigsaw puzzles (Noroozi & Favaro, 2016; Xu et al., 2019) have been recently adapted to facilitate reinforcement learning with verifiable rewards (RLVR) in MLLMs (Zeng et al., 2025; Wu et al., 2025c), relying on a single pretext task limits generalization. Recent efforts broaden this paradigm to provide annotation-free rewards: Spatial-SSRL (Liu et al., 2026) leverages 2D and 3D spatial pretexts, while VideoSSR (He et al., 2026) incorporates anomaly grounding, counting, and temporal jigsaw to mitigate annotation bias. However, existing pretexts either stay within a single visual stream or supervise only segment-level ordering, without enforcing onset-precise cross-modal correspondence. We instead insert paired audio-visual events at shared onsets, providing supervision that directly targets fine-grained temporal binding.

## 3 Method

### 3.1 Overview

We observe that current audio–visual models often struggle with temporal alignment, including inaccurate localization and weak cross-modal correspondence. To directly target this ability, we propose an alignment-focused self-supervised RL framework that inserts synthetic audio–visual events at controlled timestamps as supervision, and trains the model on the resulting QA instances with GRPO under an alignment-aware reward.

Although built from synthetic events, our design targets a capability rather than a data distribution. Pure tones and geometric shapes are used precisely because their recognition is trivial for pretrained encoders, so the reward gradient concentrates on when the two streams should be bound rather than what they contain. Under GRPO, the predicted timestamp further acts as an intermediate reasoning step for cross-modal querying, encouraging a general binding procedure rather than surface mappings from event identity to time. Since this procedure is agnostic to the event vocabulary, it is expected to transfer to real-world audio–visual reasoning, where semantic content differs but temporal binding remains a shared bottleneck. We next describe how this design is instantiated as a dataset (Section 3.2) and a training procedure (Section 3.3).

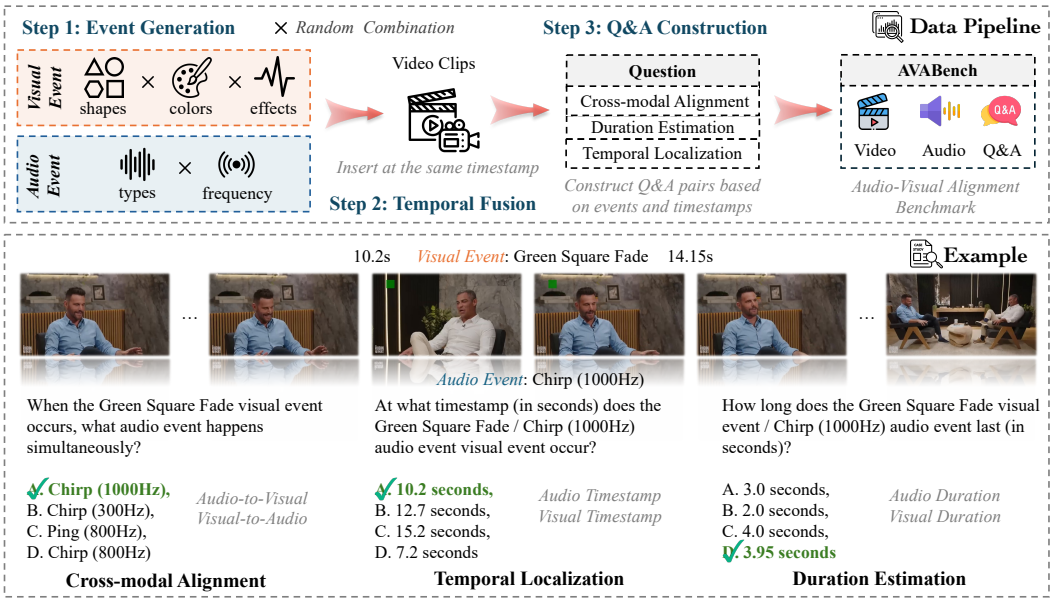


Figure 2: **Overview of AVABench.** Visual and audio events are randomly combined and temporally fused into video clips by inserting them at same timestamps, from which question-answer pairs are constructed to evaluate cross-modal alignment, temporal localization, and duration estimation.

### 3.2 Constructed Alignment Dataset

We build our alignment dataset on top of OmniVideo. For each video $v$ we sample a contiguous clip $x_v$ and insert a synthetic visual event and an audio event that share the same onset timestamp.

**Task Design.** We introduce a constructed Audio-Visual Temporal Alignment QA task that tests whether a model can jointly localize an event in time and verify its cross-modal correspondence between the visual and audio streams. Given a video clip $x'_v$ and an associated multiple-choice question $q$, the model must select the option that correctly identifies (i) the onset timestamp at which a paired audio-visual event occurs and (ii) the matching relationship between the inserted visual event identity and the inserted audio event identity. Unlike conventional audio-visual QA benchmarks, which mostly probe coarse-grained semantic consistency (e.g., "does this sound match this scene"), our task requires fine-grained temporal grounding: the model cannot rely on holistic scene semantics alone but must precisely attend to when the coupled event happens and whether the two modalities are genuinely synchronized rather than merely co-occurring somewhere in the clip. This makes the task a direct probe of a model's underlying audio-visual synchrony representation, a capability that is essential for downstream applications such as lip-sync and dubbing quality verification, deepfake/video-tampering detection, precise audio-visual event retrieval, and surveillance or sports analytics where the exact timing of a cross-modal event matters. Because the supervision signal is derived purely from controlled insertion metadata rather than human annotation, the task also provides a scalable and leakage-resistant way to benchmark and further fine-tune multimodal models on temporal alignment ability.

**Dataset Construction.** To construct each synthetic training/evaluation instance, we sample an insertion time $t$ uniformly from the valid range of a source video $x_v$, and choose a visual event identity $u$ and an audio event identity $a$ from fixed vocabularies. Specifically, each visual event $u$ is defined by a combination of attributes drawn from the following vocabularies:

- **Color:** red, blue, green, yellow, purple, orange
- **Shape:** circle (•), square (■), triangle (▲), star (★), diamond (♦), cross (+)
- **Effect:** flash, pulse, blink, fade, grow, rotate

and each audio event $a$ is defined by a combination of attributes drawn from:

- **Audio type:** beep, boop, chirp, burst, sweep, ping
- **Frequency:** 300, 500, 800, 1000, 1500, 2000 Hz

This yields a combinatorially large space of distinct visual and audio event identities ($6\times6\times6 = 216$ visual event types and $6 \times 6 = 36$ audio event types), which reduces the chance of the model memorizing a small fixed set of event patterns and encourages generalization over compositional attribute combinations. We inject the visual event with onset $t$ and duration $d$, and inject the audio event at the same onset $t$:

$$V(x_v, t, u, d) = \text{InsertVisual}(x_v, t, u, d), \qquad A(x_v, t, a) = \text{InsertAudio}(x_v, t, a).$$

The final synthetic clip is obtained by mixing the original video with the inserted visual and audio events,

$$x'_v = \text{Mix}(x_v,\ V(x_v, t, u, d),\ A(x_v, t, a)),$$

so that by construction the temporal coupling $t^{(v)}_{\text{onset}} = t^{(a)}_{\text{onset}} = t$ holds exactly. For every inserted event we then generate a multiple-choice question $q$ covering temporal localization (identifying $t$) and cross-modal alignment (identifying the correct $(u, a)$ correspondence). The model is given $(x'_v, q)$ and must select the option matching both the correct onset and the correct audio-visual correspondence; supervision targets are read off directly from the insertion metadata $(t, d, u, a)$ without any human annotation. To ensure reliable evaluation, we enforce strict video-level non-overlap between the training and test splits, preventing the same source video from appearing in both.

### 3.3 Self-Supervised RL

We train the model with self supervised RL to improve audio visual temporal alignment. For each input video clip $v$ and query $q$ the policy $\pi_\theta$ generates a structured response

$$o = \langle\texttt{timestamp}\rangle\langle\texttt{thinking}\rangle\langle\texttt{answer}\rangle$$

where the timestamp is the predicted onset time and the answer selects the correct option.

We optimize the expected reward

$$J(\theta) = \mathbb{E}_{o\sim\pi_\theta(\cdot|v,q)}\left[R(o)\right]. \tag{1}$$

**Reward Design.** We use a composite reward with three parts. **Timestamp reward** measures the absolute error between the predicted timestamp and the ground truth onset timestamp from our insertion process with stepwise values. **Format reward** is 1 if the output follows the required timestamp thinking answer structure otherwise 0. **Accuracy reward** is 1 if the chosen option is correct otherwise 0. The final reward is

$$R(o) = 0.1\, r_{\text{time}} + 0.1\, r_{\text{fmt}} + 0.8\, r_{\text{acc}}. \tag{2}$$

**Optimization.** We adopt GRPO to optimize the policy without the memory overhead of a value network. For each query, we sample a group of $G$ outputs from the current policy and estimate the baseline using the group average. The surrogate loss is

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\left[\min\Big(r_i(\theta)A_i,\ \text{clip}\big(r_i(\theta), 1-\epsilon, 1+\epsilon\big)A_i\Big) - \beta D_{KL}\right], \tag{3}$$

where $r_i(\theta) = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}$ is the probability ratio, $\epsilon$ is the clipping parameter, and $\beta$ controls the KL penalty against the reference policy.

## 4 Experiments

### 4.1 Experimental Setup

**Benchmarks and Datasets** We evaluate our method on our constructed AVABench and four established audio-visual benchmarks to assess both effectiveness and generalization. Specifically, we

Table 1: **Experimental Results on OmniVideo-100K.** Performance comparison of open-source multimodal LLMs of varying sizes on OmniVideo-100K, covering Alignment, Understanding, and Reasoning, as well as two video duration ranges. Our model consistently outperforms the similarly sized Qwen2.5-Omni-3B baseline across all metrics.

| Models | Size | Alignment | Understanding | Reasoning | (0, 2]min | (2, 5]min | Avg |
|---|---|---|---|---|---|---|---|
| *Open-source MLLMs* | | | | | | | |
| VITA-1.5 (Fu et al., 2025b) | 7B | 28.45 | 43.02 | 48.09 | 44.72 | 39.24 | 40.99 |
| video-SALMONN 2+ (Tang et al., 2025) | 7B | 37.93 | 49.22 | 43.51 | 46.58 | 44.48 | 45.15 |
| Qwen2.5-Omni-7B (Xu et al., 2025a) | 7B | 39.66 | 46.51 | 38.17 | 38.51 | 44.77 | 42.77 |
| Uni-MoE-2.0-Omni (Li et al., 2025c) | 30B | 39.66 | 52.71 | 41.98 | 46.58 | 47.09 | 46.93 |
| Qwen3-Omni (Xu et al., 2025b) | 30B | 43.10 | 55.04 | 45.04 | 49.07 | 50.00 | 49.70 |
| OmniVinci (Ye et al., 2025a) | 9B | 43.97 | 51.55 | 41.22 | 47.20 | 47.09 | 47.13 |
| MiniCPM-o 4.5 (Yao et al., 2024) | 9B | 56.90 | 58.14 | 48.09 | 53.42 | 56.10 | 55.25 |
| Qwen2.5-Omni-3B (Xu et al., 2025a) | 3B | 33.62 | 45.35 | 40.46 | 47.20 | 38.66 | 41.39 |
| **Omni-Aligner (Ours)** | 3B | 37.07 (+3.45) | 49.22 (+3.87) | 45.04 (+4.58) | 50.93 (+3.73) | 42.73 (+4.07) | 45.35 (+3.96) |

first test on AVABench, our self-constructed dataset. Then we further evaluate on: (1) OmniVideo-100K (Cai et al., 2026): an audio-visual instruction-tuning benchmark with evidence-grounded, entity-consistent long-range reasoning; (2) WorldSense (Hong et al., 2025): an omni-modal video understanding benchmark with synchronized audio-visual inputs and high-quality multiple-choice QA requiring true cross-modal integration; (3) OmniVideoBench (Li et al., 2025a): a reasoning-focused benchmark with verified QA and explicit step-by-step traces over diverse video durations; and (4) Daily-Omni (Zhou et al., 2025): a multiple-choice audio-visual QA benchmark that emphasizes temporal alignment across modalities on real-world 30s/60s clips.

**Implementation Details** We fine-tune Qwen2.5-Omni-3B via GRPO on 8 NVIDIA A100 GPUs with DeepSpeed ZeRO-3 and `bf16` precision. FlashAttention-2 and gradient checkpointing are enabled. For each prompt, $G = 8$ completions are sampled at temperature 0.7 (`top-p` $= 0.95$, `top-k` $= 50$). The model is trained for 1 epoch with a learning rate of $1 \times 10^{-6}$, cosine schedule, weight decay 0.01, max gradient norm 5.0, and KL coefficient $\beta = 0.001$. The maximum prompt and completion lengths are 8,192 and 768 tokens, respectively. Video frames are extracted at 0.5 FPS with a pixel budget of 112,896. More details can be found in the appendix.

### 4.2 Results on Constructed Alignment Dataset

Table 2 shows that existing audio-visual temporal alignment models perform poorly on our dataset despite the task being relatively straightforward. For instance, even the strongest reported baseline, Qwen2.5-Omni-7B, reaches only 50.86 overall, and most models stay far from accurate temporal alignment, suggesting a clear weakness in learning reliable cross-modal temporal correspondence. The results are also uneven across sub-tasks, with noticeable degradation on cross-modal alignment and time reasoning, such as A→V at 42.46 and CMA around the mid 50s for the best baseline. In contrast, our trained model, Omni-Aligner, achieves SOTA across all reported categories, reaching 74.77 overall. It improves consistently on temporal attribute estimation such as AD, AT, VD, and VT, as well as on alignment and localization such as DE and TL, where it attains 78.67 and 86.90 respectively, demonstrating that our method effectively addresses the core limitations of prior approaches.

### 4.3 Results on Current Audio-Visual Understanding Benchmarks

Across OmniVideo-100K and WorldSense in Tables 1 and 3, our model, sharing the same three-billion-parameter backbone as Qwen2.5-Omni-3B, consistently outperforms this baseline. On OmniVideo-100K, it improves Alignment by 3.45 points, Understanding by 3.87 points, and Reasoning by 4.58 points, with gains of 3.73 and 4.07 points on short and longer videos respectively, for an average improvement of 3.96 points. On WorldSense, our model improves over the backbone in all eight domains, raising the average from 41.4 to 42.9, with the largest gains in Culture and Politics and in Games. Notably, at this scale our 3B model surpasses several larger open-source MLLMs, including the 7B VITA-1.5, 8B InternVL2.5, and 7B LLaVA-Video, and edges above GPT-4o on average, 42.9 versus 42.6.

Table 2: Comparison of different methods on alignment accuracy. Abbreviations in the table are defined as follows: AD = Audio Duration, AT = Audio Timestamp, A→V = Cross-modal Audio-to-Visual, V→A = Cross-modal Visual-to-Audio, VD = Visual Duration, VT = Visual Timestamp, CMA = Cross-modal Alignment, DE = Duration Estimation, and TL = Temporal Localization.

| Method | Size | Task Type | | | | | | Question Category | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AD | AT | A→V | V→A | VD | VT | CMA | DE | TL | |
| *Proprietary Models* | | | | | | | | | | | |
| Gemini 3.1 Pro (Google DeepMind, 2026) | - | 65.48 | 51.59 | 62.50 | **75.40** | 37.70 | 81.55 | **69.94** | 61.51 | 65.58 | 65.65 |
| *Open-source MLLMs* | | | | | | | | | | | |
| Video-SALMONN 2+ (Tang et al., 2025) | 7B | 26.79 | 17.46 | 33.73 | 18.85 | 10.32 | 26.98 | 22.81 | 27.38 | 16.87 | 22.36 |
| VITA-1.5 (Fu et al., 2025b) | 7B | 27.78 | 20.83 | 32.74 | 23.81 | 15.87 | 29.76 | 26.79 | 25.79 | 22.82 | 25.17 |
| OmniVinci (Ye et al., 2025a) | 9B | 29.56 | 23.02 | 33.73 | 20.83 | 22.22 | 24.60 | 29.37 | 25.00 | 22.62 | 25.66 |
| MiniCPM-o 4.5 (Yao et al., 2024) | 9B | 33.93 | 30.95 | 35.91 | 39.88 | 25.60 | 25.79 | 42.06 | 30.95 | 23.02 | 32.01 |
| Uni-MoE-2.0-Omni (Li et al., 2025c) | 30B | 33.93 | 34.52 | 38.29 | 41.67 | 29.96 | 22.42 | 42.46 | 31.65 | 28.27 | 34.10 |
| Qwen2.5-Omni-3B (Xu et al., 2025a) | 3B | 41.07 | 45.44 | 36.31 | 42.46 | 39.88 | 29.76 | 39.38 | 40.48 | 37.60 | 39.15 |
| Qwen2.5-Omni-7B (Xu et al., 2025a) | 7B | 44.44 | 57.54 | 42.46 | 69.05 | 45.24 | 46.43 | 55.75 | 44.84 | 51.98 | 50.86 |
| **Omni-Aligner (Ours)** | **3B** | **75.79** | **90.08** | **44.84** | 72.62 | **81.55** | **83.73** | 58.73 | **78.67** | **86.90** | **74.77** |

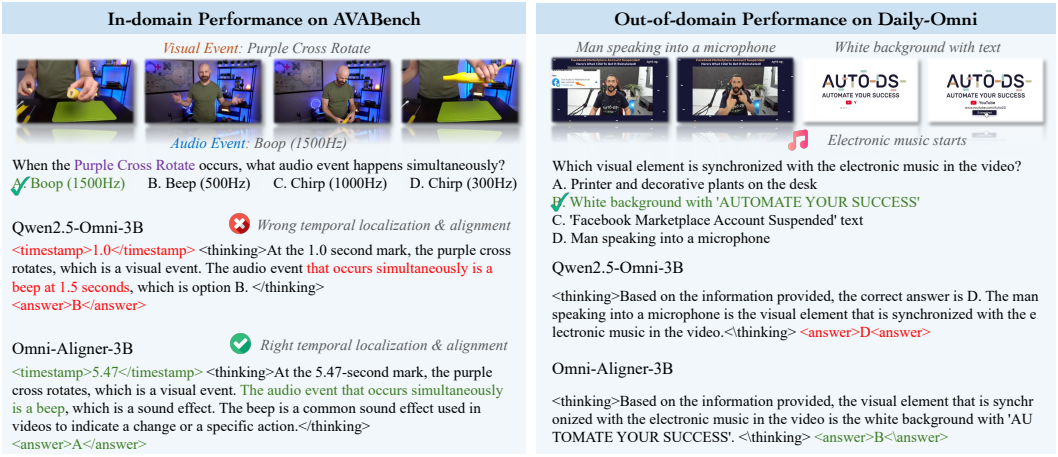


Figure 3: **Quatitative Results.** Qualitative comparison between Qwen2.5-Omni-3B and our Omni-Aligner-3B, showing in-domain results on AVABench and out-of-domain results on Daily-Omni. Our model achieves more accurate audio-visual temporal localization and alignment, leading to correct answers where the baseline fails due to misaligned event grounding.

This advantage extends to more challenging omni-modal settings. On OmniVideoBench in Table 5, our model improves the backbone's average from 29.20 to 30.00, with gains across all three audio types and most video-duration ranges, though the improvement is largest for videos under one minute and narrows slightly on longer clips, with a minor drop on the 5 to 10 minute range. On Daily-Omni in Table 6, our model improves five of six sub-dimensions, most notably Inference by 3.90 points and Context Understanding by 3.10 points, raising the overall average by 2.00 points. Taken together, these consistent, broadly distributed improvements across alignment, understanding, and reasoning, and across varying audio types and video durations, indicate that our training approach yields robust gains over the same-sized backbone, with only minor variation across longer temporal ranges pointing to room for further improvement.

### 4.4 Qualitative Results

Fig. 3 presents representative cases comparing Qwen2.5-Omni-3B and our Omni-Aligner-3B. On the in-domain AVABench benchmark, given a video containing the visual event Purple Cross Rotate and the task of identifying its synchronized audio event, the baseline mislocalizes the event timestamp at 1.0s and consequently selects the wrong audio event, Beep. In contrast, Omni-Aligner-3B

Table 3: **Performance comparison on Daily-Omni and WorldSense.** We report three representative sub-categories along with the overall average for each benchmark. AV Align = audio-visual alignment, Ctx. Und. = context understanding, Infer. = inference. Complete results are provided in the appendix.

| Methods | Daily-Omni | | | | WorldSense | | | |
|---|---|---|---|---|---|---|---|---|
| | AV Align | Ctx. Und. | Infer. | Avg | Games | Sports | Music | Avg |
| *Open-Source Omni-Modal / Video-Audio MLLMs* | | | | | | | | |
| OneLLM (Han et al., 2024) | – | – | – | – | 23.7 | 22.4 | 19.8 | 22.8 |
| Unified-IO-2 L (Lu et al., 2024) | 27.31 | 26.42 | 29.87 | 27.40 | 24.1 | 22.9 | 25.3 | 23.3 |
| Unified-IO-2 XL (Lu et al., 2024) | 30.25 | 25.39 | 33.12 | 28.32 | 28.0 | 25.7 | 24.2 | 24.7 |
| Unified-IO-2 XXL (Lu et al., 2024) | 25.63 | 26.42 | 35.06 | 28.24 | 23.7 | 25.7 | 27.3 | 25.9 |
| VideoLLaMA2 (Cheng et al., 2024) | 35.71 | 35.75 | 40.91 | 35.17 | 24.6 | 25.5 | 27.1 | 25.4 |
| VITA-1.5 (Fu et al., 2025a) | – | – | – | – | 32.6 | 34.7 | 39.9 | 36.9 |
| Ola (Liu et al., 2025c) | 40.34 | 40.41 | 63.64 | 50.71 | – | – | – | – |
| Qwen2.5-Omni-7B (Xu et al., 2025a) | 48.32 | **58.55** | 76.62 | 62.07 | – | – | – | – |
| Qwen2.5-Omni-3B (Xu et al., 2025a) | 50.84 | 53.89 | 75.97 | 60.23 | 34.7 | 39.5 | 40.1 | 41.4 |
| **Omni-Aligner (Ours)** | **52.94** (+2.10) | 56.99 (+3.10) | **79.87** (+3.90) | **62.23** (+2.00) | **36.9** (+2.2) | **40.0** (+0.5) | **41.4** (+1.3) | **42.9** (+1.5) |

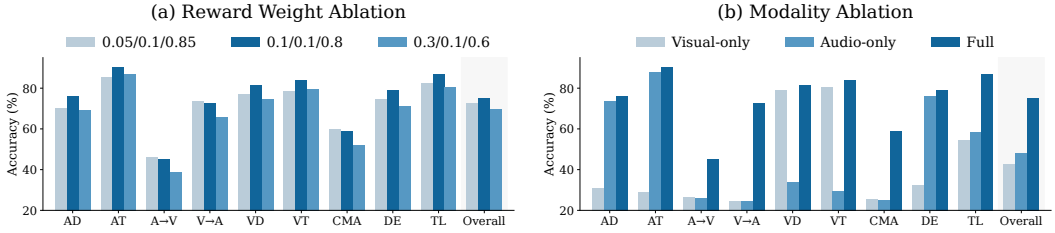


Figure 4: Ablation studies on (a) reward weight configurations and (b) training modality combinations, evaluated across all AVABench sub-categories.

accurately grounds the event at 5.47s and correctly identifies the co-occurring audio event, demonstrating that our alignment training substantially improves fine-grained temporal localization.

On the more challenging out-of-domain Daily-Omni benchmark, the baseline is misled by visual saliency, incorrectly associating the onset of electronic music with the most prominent visual element, a man speaking, rather than the actually synchronized event. Omni-Aligner-3B correctly identifies that the background text change reading AUTOMATE YOUR SUCCESS is the visual event truly synchronized with the music. These cases show that the temporal alignment ability learned by our model generalizes well beyond the training distribution, enabling reasoning grounded in genuine audio-visual synchrony rather than superficial visual cues.

### 4.5 Ablation Study

We conduct three ablation studies on AVABench and OmniVideo-100K to analyze the impact of training data, reward design, and optimization paradigm on Omni-Aligner.

**Training Subset Ablation.** As shown in Table 4a, the full subset achieves the best performance (74.77 on AVABench, 45.35 on OmniVideo-100K), while single-modality or partial supervision is generally less effective. Alignment and temporal localization subsets yield the strongest gains on AVABench, indicating the importance of grounding-related signals.

**Reward Design Ablation.** Table 4b shows that removing any of the timestamp, format, or accuracy reward degrades AVABench performance, with the full reward achieving the best score. Figure 4a further compares three reward weight configurations, where the balanced 0.1/0.1/0.8 setting performs best, indicating that overemphasizing any single reward hurts joint optimization. Figure 4b shows that the Full audio-visual setting consistently outperforms single-modality variants, especially on A→V, V→A, and CMA, confirming that joint supervision is essential for learning genuine cross-modal alignment rather than uni-modal shortcuts.

**Optimization Paradigm Ablation.** Figure 5 compares different optimization paradigms across AVABench and four out-of-domain benchmarks. Pure GRPO consistently achieves the best perfor-

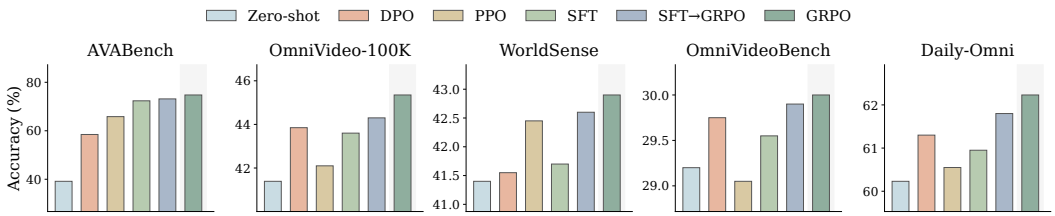


Figure 5: Comparison of different post-training strategies (Zero-shot, DPO, PPO, SFT, SFT→GRPO, and GRPO) on AVABench and four out-of-domain audio-visual benchmarks.

Table 4: **Ablation studies.** We evaluate the contribution of different training elements.

(a) **Ablation on Different Training Subsets.**

| Trained Subset | AVABench | OmniVideo-100K |
|---|---|---|
| *Baseline* | | |
| Qwen2.5-Omni-3B | 39.15 | 41.39 |
| *Checkpoint Trained with Different Subsets* | | |
| Audio-only | 59.62 | 42.08 |
| Visual-only | 52.37 | 44.01 |
| Alignment | 68.44 | 43.17 |
| Temporal Localization | 61.93 | 42.62 |
| Full subset | **74.77** | **45.35** |

(b) **Ablation on GRPO Reward Design.**

| Reward Setting | AVABench | OmniVideo-100K |
|---|---|---|
| *Baseline* | | |
| Qwen2.5-Omni-3B | 39.15 | 41.39 |
| *Reward component ablation (GRPO with different rewards).* | | |
| w/o timestamp ($r_{\text{time}} = 0$) | 71.08 | 44.26 |
| w/o format ($r_{\text{fmt}} = 0$) | 55.41 | 42.34 |
| w/o accuracy ($r_{\text{acc}} = 0$) | 64.92 | 43.52 |
| Full (time + fmt + acc) | **74.77** | **45.35** |

mance, substantially surpassing SFT and preference-based methods such as DPO and PPO. Notably, SFT→GRPO underperforms pure GRPO, suggesting that token-level imitation via SFT constrains exploration and limits the generalization benefits of reward-driven RL, supporting our choice of a purely self-supervised RL framework.

## 5 Limitations

While our framework demonstrates improved temporal alignment, several directions warrant further investigation. First, AVABench is intended as a controlled proxy task rather than a direct simulation of real-world audio-visual asynchrony. Its synthetic event vocabulary enables scalable, annotation-free supervision at low cost, whereas collecting comparable real-world data would require substantially more effort, so the current design is a preliminary step rather than a final solution. The consistent gains across four out-of-domain benchmarks suggest this simplified setting already yields transferable capability, and future work will extend toward more diverse and naturalistic event combinations to further close the gap with real-world scenarios. Second, extending the reward mechanism beyond point-based onset/duration alignment to handle overlapping events and long-range temporal dependencies would enhance robustness in intricate narratives. Finally, exploring the interplay between model scale and explicit alignment training remains important for optimizing the efficiency–performance trade-off in foundation models.

## 6 Conclusions

In this work, we identified temporal misalignment as a critical bottleneck in audio-visual understanding and proposed Omni-Aligner, a self-supervised reinforcement learning framework, alongside AVABench, an annotation-free benchmark for rigorous evaluation. By leveraging synthetic event insertion and GRPO optimization, our approach effectively enhances cross-modal temporal correspondence without requiring external annotations or models. Empirical results show that Omni-Aligner achieves state-of-the-art performance on AVABench and consistently improves upon baseline backbones across multiple out-of-domain benchmarks. These findings suggest that explicitly addressing temporal misalignment contributes significantly to the reliability and generalization of omni-modal video understanding systems.

### AI USE STATEMENT

We used generative AI tools only to edit the manuscript for readability (e.g., grammar and phrasing). No other required or recommended disclosure items apply to this work. All AI-assisted text was reviewed by the authors, who take full responsibility for the final content of this paper.

## REFERENCES

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7558–7567, 2019.

Anthropic. Introducing the next generation of Claude. https://www.anthropic.com/news/claude-3-family, 2024. Accessed: 2024-10-22.

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.

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

Jianghan Chao, Jianzhang Gao, Wenhui Tan, Yuchong Sun, Ruihua Song, and Liyun Ru. Jointavbench: A benchmark for joint audio-visual reasoning evaluation. *arXiv preprint arXiv:2512.12772*, 2025.

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

Zhangquan Chen, Jiale Tao, Ruihuang Li, Yihao Hu, Ruitao Chen, Zhantao Yang, Xinlei Yu, Haodong Jing, Manyuan Zhang, Shuai Shao, et al. Omnivideo-r1: Reinforcing audio-visual reasoning with query intention and modality attention. *arXiv preprint arXiv:2602.05847*, 2026b.

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

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

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

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025a.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025b.

Honghao Fu, Yuan Ouyang, Kai-Wei Chang, Yiwei Wang, Zi Huang, and Yujun Cai. Contextnav: Towards agentic multimodal in-context learning. In *International Conference on Learning Representations*, volume 2026, pp. 74934–74959, 2026a.

Honghao Fu, Miao Xu, Yiwei Wang, Dailing Zhang, Jun Liu, and Yujun Cai. Videostir: Understanding long videos via spatio-temporally structured and intent-aware rag. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35793–35805, 2026b.

Haonan Ge, Yiwei Wang, Kai-Wei Chang, Hang Wu, and Yujun Cai. Framemind: Frame-interleaved video reasoning via reinforcement learning. *arXiv preprint arXiv:2509.24008*, 2025.

Haonan Ge, Yiwei Wang, Hang Wu, and Yujun Cai. What should a streaming video model remember? *arXiv preprint arXiv:2606.16353*, 2026.

Tiantian Geng, Teng Wang, Jinming Duan, Yanfu Zhang, Weili Guan, Feng Zheng, and Ling Shao. Uniav: Unified audio-visual perception for multi-task video event localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11):10280–10294, 2025a. doi: 10.1109/TPAMI.2025.3593932.

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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18959–18969, 2025b.

Google DeepMind. Gemini 3.1 Pro model card. `https://deepmind.google/models/model-cards/gemini-3-1-pro/`, 2026.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26584–26595, 2024.

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. Videossr: Video self-supervised reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26530–26540, 2026.

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Yiduo Jia, Muzhi Zhu, Hao Zhong, Mingyu Liu, Yuling Xi, Hao Chen, Bin Qin, Yongjie Yang, Zhenbo Luo, and Chunhua Shen. Omnijigsaw: Enhancing omni-modal reasoning via modality-orchestrated reordering. *arXiv preprint arXiv:2604.08209*, 2026.

Mohammed Irfan Kurpath, Jaseel Muhammad Kaithakkodan, Jinxing Zhou, Sahal Shaji Mullappilly, Mohammad Almansoori, Noor Ahsan, Beknur Kalmakhanbet, Sambal Shikhar, Rishabh Lalla, Jean Lahoud, et al. A benchmark and agentic framework for omni-modal reasoning and tool use in long videos. *arXiv preprint arXiv:2512.16978*, 2025.

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

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

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19108–19118, 2022.

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

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

Zinuo Li, Xian Zhang, Yongxin Guo, Mohammed Bennamoun, Farid Boussaid, Girish Dwivedi, Luqi Gong, and Qiuhong Ke. Watch and listen: Understanding audio-visual-speech moments with multimodal llm. *arXiv preprint arXiv:2505.18110*, 2025d.

Chang Liu, Hongkai Chen, Yujun Cai, Hang Wu, Qingwen Ye, Ming-Hsuan Yang, and Yiwei Wang. Structured attention matters to multimodal llms in document understanding. *arXiv preprint arXiv:2506.21600*, 2025a.

Kai Liu, Jungang Li, Yuchong Sun, Shengqiong Wu, Jianzhang Gao, Daoan Zhang, Wei Zhang, Sheng Jin, Sicheng Yu, Geng Zhan, et al. Javisgpt: A unified multi-modal llm for sounding-video comprehension and generation. *arXiv preprint arXiv:2512.22905*, 2025b.

Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. Spatial-ssrl: Enhancing spatial understanding via self-supervised reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9570–9581, 2026.

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

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26439–26455, 2024.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pp. 69–84. Springer, 2016.

Madeline C Schiappa, Yogesh S Rawat, and Mubarak Shah. Self-supervised learning for videos: A survey. *ACM Computing Surveys*, 55(13s):1–37, 2023.

Bowen Sun, Yujun Cai, Ming-Hsuan Yang, Hang Wu, and Yiwei Wang. Pas: A training-free stabilizer for temporal encoding in video llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14471–14480, 2026.

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.

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.

Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *European conference on computer vision*, pp. 504–521. Springer, 2020.

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

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

Hang Wu, Yujun Cai, Haonan Ge, Hongkai Chen, Ming-Hsuan Yang, and Yiwei Wang. Refineshot: Rethinking cinematography understanding with foundational skill evaluation. *arXiv preprint arXiv:2510.02423*, 2025a.

Hang Wu, Hongkai Chen, Yujun Cai, Chang Liu, Qingwen Ye, Ming-Hsuan Yang, and Yiwei Wang. Dimo-gui: Advancing test-time scaling in gui grounding via modality-aware visual reasoning. In *EMNLP 2025*, 2025b.

Hang Wu, Yujun Cai, Zehao Li, Haonan Ge, Bowen Sun, Junsong Yuan, and Yiwei Wang. Camreasoner: Reinforcing camera movement understanding via structured spatial reasoning. *arXiv preprint arXiv:2602.00181*, 2026a.

Hang Wu, Sherin Mary Mathews, Yujun Cai, Ming-Hsuan Yang, and Yiwei Wang. Semantic-aware adaptive visual memory for streaming video understanding. *arXiv preprint arXiv:2605.07897*, 2026b.

Penghao Wu, Yushan Zhang, Haiwen Diao, Bo Li, Lewei Lu, and Ziwei Liu. Visual jigsaw post-training improves mllms. *arXiv preprint arXiv:2509.25190*, 2025c.

Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10334–10343, 2019.

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, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 3480–3491, 2022.

Zhihao Yang, Ancheng Xu, Jingpeng Li, Liang Yan, Jiehui Zhou, Zhen Qin, Hengyu Chang, Yukun Chen, Longze Chen, Ahmadreza Argha, et al. Structuring reasoning for complex rules beyond flat representations. *arXiv preprint arXiv:2510.05134*, 2025.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv preprint arXiv:2510.15870*, 2025a.

Qilang Ye, Zitong Yu, Rui Shao, Xinyu Xie, Philip Torr, and Xiaochun Cao. Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios. In *European Conference on Computer Vision*, pp. 146–164. Springer, 2024.

Qilang Ye, Zitong Yu, Rui Shao, Yawen Cui, Xiangui Kang, Xin Liu, Philip Torr, and Xiaochun Cao. Cat+: Investigating and enhancing audio-visual understanding in large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(10):8674–8690, 2025b. doi: 10.1109/TPAMI.2025.3582389.

Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2031–2041, 2021.

Table 5: Performance comparison on OmniVideoBench across baseline LLMs, visual-only VLMs, and omni-modal language models, evaluated across audio types and video durations. Our model improves over the similarly sized Qwen2.5-Omni-3B baseline, achieving a 0.80-point gain in average accuracy.

| Models | Audio Type | | | Video Duration | | | | Avg. |
|---|---|---|---|---|---|---|---|---|
| | Music | Sound | Speech | (0,1] min | (1,5] min | (5,10] min | (10,30] min | |
| *Omni-Modal Language Models (With Visual and Audio)* | | | | | | | | |
| Qwen3-Omni-30B-A3B | 37.36 | 34.67 | 39.26 | 45.78 | 37.03 | 38.86 | 35.11 | 38.40 |
| OmniVinci-9B | 30.77 | 32.67 | 32.15 | 38.55 | 34.11 | 30.13 | 27.10 | 32.10 |
| MiniCPM-o | 27.47 | 28.57 | 30.24 | 31.43 | 28.49 | 34.53 | 26.15 | 29.70 |
| Qwen2.5-Omni-7B | 23.07 | 25.33 | 30.70 | 41.57 | 27.41 | 25.33 | 26.72 | 29.30 |
| VideoLLaMA2-7B | 26.37 | 30.67 | 29.25 | 32.00 | 28.20 | 29.60 | 28.29 | 29.20 |
| Qwen2.5-Omni-3B | 24.18 | 25.17 | 30.58 | 29.82 | 28.11 | 30.13 | 29.39 | 29.20 |
| **Omni-Aligner (Ours)** | 26.37 | 27.89 | 30.84 | 34.50 | 28.40 | 29.26 | 29.96 | 30.00 (+0.80) |

Table 6: **Performance comparison of MLLMs on Daily-Omni**. Boldface and underline indicate the top two performers. Abbreviations: AV Align = audio-visual alignment, Comp. = comparative, Ctx. Und. = context understanding, Evt. Seq. = event sequence, Infer. = inference, Reas. = reasoning.

| Methods | AV Align | Comp. | Ctx. Und. | Evt. Seq. | Infer. | Reas. | Avg |
|---|---|---|---|---|---|---|---|
| *Omni-Modal Language Models (With Visual and Audio)* | | | | | | | |
| Qwen2.5-Omni-7B | 48.32 | 69.47 | 58.55 | 58.17 | 76.62 | 73.14 | 62.07 |
| Ola (7B) | 40.34 | 61.07 | 40.41 | 43.46 | 63.64 | 69.71 | 50.71 |
| Unified-IO-2 L (1B) | 27.31 | 22.90 | 26.42 | 27.78 | 29.87 | 29.14 | 27.40 |
| Unified-IO-2 XL (3B) | 30.25 | 30.53 | 25.39 | 29.08 | 33.12 | 21.71 | 28.32 |
| Unified-IO-2 XXL (8B) | 25.63 | 31.30 | 26.42 | 25.82 | 35.06 | 29.71 | 28.24 |
| VideoLLaMA2 (7B) | 35.71 | 35.88 | 35.75 | 31.70 | 40.91 | 34.29 | 35.17 |
| Qwen2.5-Omni-3B | 50.84 | 69.47 | 53.89 | 53.92 | 75.97 | 70.29 | 60.23 |
| **Omni-Aligner (Ours)** | 52.94 (+2.10) | 70.99 (+1.52) | 56.99 (+3.10) | 55.56 (+1.64) | 79.87 (+3.90) | 70.29 (-) | 62.23 (+2.00) |

Yu Zeng, Wenxuan Huang, Shiting Huang, Xikun Bao, Yukun Qi, Yiming Zhao, Qiuchen Wang, Lin Chen, Zehui Chen, Huaian Chen, et al. Agentic jigsaw interaction learning for enhancing visual perception and reasoning in vision-language models. *arXiv preprint arXiv:2510.01304*, 2025.

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

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

# A Appendix

## A.1 More Quantitative Results

## A.2 More Implementation Details

Table 8 summarizes the complete set of hyperparameters used during the GRPO-based alignment stage of EchoInk. The base model is Qwen2.5-Omni-3B, fine-tuned in full-parameter mode (no LoRA/PEFT) on 8 NVIDIA GPUs with bf16 mixed precision. The reward signal combines format compliance, answer accuracy, and timestamp fidelity via a weighted sum, while a small KL penalty ($\beta = 0.001$) keeps the policy close to the reference model.

Table 7: Performance comparison on WorldSense across proprietary, open-source video-audio, and open-source video MLLMs of varying sizes and domains. Our model improves over the similarly sized Qwen2.5-Omni-3B baseline across all domains, with a 1.5-point gain on average.

| Methods | LLM Size | Tech & Science | Culture & Politics | Daily Life | Film & TV | Performance | Games | Sports | Music | Avg |
|---|---|---|---|---|---|---|---|---|---|---|
| *Proprietary MLLMs* | | | | | | | | | | |
| Claude 3.5 Sonnet (Anthropic, 2024) | - | 43.7 | 31.7 | 30.6 | 36.5 | 30.7 | 31.9 | 36.6 | 33.9 | 34.8 |
| GPT 4o (Hurst et al., 2024) | - | 48.0 | 44.0 | 38.3 | 43.5 | 41.9 | 41.2 | 42.6 | 42.7 | 42.6 |
| *Open-Source Video-Audio MLLMs* | | | | | | | | | | |
| OneLLM (Han et al., 2024) | 7B | 26.7 | 25.1 | 19.0 | 22.7 | 27.0 | 23.7 | 22.4 | 19.8 | 22.8 |
| Unified-IO-2 L (Lu et al., 2024) | 1B | 19.3 | 22.8 | 23.1 | 25.6 | 25.8 | 24.1 | 22.9 | 25.3 | 23.3 |
| Unified-IO-2 XL (Lu et al., 2024) | 3B | 26.5 | 24.4 | 22.5 | 23.5 | 24.7 | 28.0 | 25.7 | 24.2 | 24.7 |
| VideoLLaMA2 (Cheng et al., 2024) | 7B | 29.4 | 25.4 | 21.8 | 24.5 | 26.2 | 24.6 | 25.5 | 27.1 | 25.4 |
| Unified-IO-2 XXL (Lu et al., 2024) | 7B | 27.1 | 31.7 | 23.9 | 23.7 | 25.5 | 23.7 | 25.7 | 27.3 | 25.9 |
| VITA-1.5 (Fu et al., 2025a) | 7B | 38.2 | 35.9 | 34.3 | 39.8 | 41.2 | 32.6 | 34.7 | 39.9 | 36.9 |
| Qwen2.5-Omni-3B | 3B | 46.5 | 43.0 | 42.1 | 44.6 | 38.9 | 34.7 | 39.5 | 40.1 | 41.4 |
| **Omni-Aligner (Ours)** | 3B | 47.3 | 46.3 | 42.4 | 45.4 | 40.8 | 36.9 | 40.0 | 41.4 | 42.9 (+1.5) |

Table 8: Alignment training hyperparameters of Omni-Aligner(GRPO, base model: Qwen2.5-Omni-3B).

| **Hyperparameter** | **Value** |
|---|---|
| *Optimization* | |
| Optimizer | AdamW |
| Learning rate | $1 \times 10^{-6}$ |
| LR scheduler | cosine |
| Weight decay | 0.01 |
| Max grad norm | 5 |
| Warmup steps | 0 |
| *Training* | |
| Training epochs | 1 |
| Per-device batch size | 1 |
| Gradient accumulation | 1 |
| Effective batch size | 8 (8 GPUs) |
| Number of GPUs | 8 (`nnodes` $=1$, NVIDIA) |
| Precision | bf16 |
| Gradient checkpointing | True |
| Loss coeff. $\beta$ (KL) | 0.001 |
| LoRA / PEFT | None (full fine-tune) |
| *GRPO* | |
| Generations per prompt | 8 |
| Reward wt. (format / ts-format) | 0.1 |
| Reward wt. (accuracy) | 0.8 |
| Reward wt. (timestamp) | 0.1 |
| Length control reward | – (disabled) |
| Weighted reward | True |
| Advantage $\varepsilon$ | $1 \times 10^{-4}$ |
| *Data / Model* | |
| Max prompt length | 8192 tokens |
| Max completion length | 768 tokens |
| Max pixels (vision) | 112,896 |
| Min pixels (vision) | 3,136 |
| Video FPS | 0.5 |
| Audio in video | True |
| Attn implementation | Flash Attention 2 |
| Eval strategy | no |
| Logging steps | 1 |
| Save steps | 1000 |