OmniSE: Support-Faithful Optimization for Evidence-Centric Audio-Video Reasoning

Xilin He[1,2], **Yifan Shen**[1], **Ankan Deria**[1], **Xiangyu Yue**[2,3], **Muhammad Haris Khan**[1]

[1]Mohamed bin Zayed University of Artificial Intelligence
[2]MMLab, The Chinese University of Hong Kong
[3]Shenzhen Loop Area Institute
xilin.he@mbzuai.ac.ae

Abstract

Omni-modal models are expected to reason over synchronized visual events and audio cues, yet answer accuracy alone can conceal failures in where an answer is grounded and which modality supports it. In audio-video question answering (AVQA), models may produce correct answers while relying on incorrect temporal spans, dominant-modality shortcuts, or unsupported cross-modal relations. We argue that this failure arises from treating reasoning as a completion-level prediction problem, where supervision and reward signals do not distinguish between support selection and reasoning over the support. To address this, we introduce OmniSE, a support-faithful optimization framework that makes question-conditioned audio-video support the central training object. Each QA instance is rewritten into an Audio-Visual Evidence Trace (AVET), a compact optimization unit that records temporal evidence, modality-specific cues, cross-modal relations, and cited reasoning steps. Given this support object, OmniSE first uses Support-Tree Rollout (STR) to explore alternative support hypotheses before sampling support-conditioned reasoning and answers. It then applies factorized reward assignment to score temporal localization, modality evidence, cross-modal binding, reasoning faithfulness, and answer correctness at their corresponding comparison levels. Finally, reward-gated self-consistency regularizes sibling continuations under reliable support, encouraging fixed trustworthy evidence to induce stable reasoning and answers. Experiments on multiple AVQA benchmarks show that OmniSE consistently outperforms strong open-source baselines at the same scale, demonstrating the importance of support-centric optimization for evidence-centric audio-video reasoning. Code and data is available at OmniSE.

1 Introduction

Audio-Visual Question Answering (AVQA) is a fundamental task where models must reason over synchronized multimodal signals. Unlike unimodal QA, AVQA requires predictions in temporally aligned audio-visual evidence, often involving both modality-specific cues and cross-modal reasoning. Reinforcement learning (RL) has recently emerged as a dominant paradigm for improving AVQA systems, particularly for enhancing reasoning quality and answer accuracy in large multimodal models [Feng et al., 2025, Li et al., 2025, Xing et al., 2025, Yang et al., 2025, Chen et al., 2026]. By optimizing generated responses with reward signals, RL encourages coherent and task-aligned outputs.

However, existing approaches remain largely answer-centric: they optimize for producing the correct final answer, without explicitly constraining whether that answer is supported in the correct audio-visual evidence [Xing et al., 2025, Liu et al., 2025]. This leads to a central limitation: correct answers

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

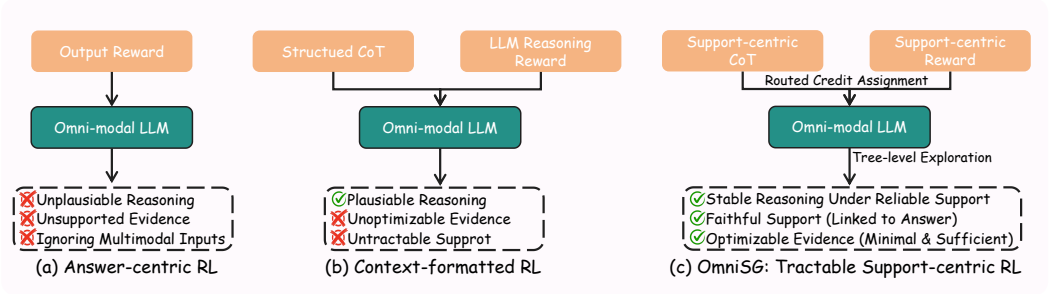


Figure 1: Comparison of different reinforcement learning paradigms for omni-modal LLMs. **(a)** Answer-centric RL [Shao et al., 2024, Xing et al., 2025, Liu et al., 2025] rewards only final outputs, leading to implausible reasoning and weak multimodal grounding. **(b)** Context-formatted RL [Yang et al., 2025, Chen et al., 2026] improves plausibility with structured CoT but yields unoptimizable evidence and untraceable support. **(c)** OmniSE, as an tracable support-centric method, assigns credit to intermediate support, enabling stable, faithful reasoning with minimal and sufficient evidence.

do not necessarily imply correct reasoning or evidence usage. In practice, a model can achieve high reward by relying on spurious correlations, dominant modalities, or irrelevant temporal regions, as long as the final answer is correct or plausible. The issue arises because current RL objectives collapse all intermediate decisions, such as evidence selection and cross-modal binding, into a single scalar reward at the end of generation, making credit assignment fundamentally ambiguous across token spans. As illustrated in Figure 1, this limitation persists across existing RL paradigms. Answer-centric RL reinforces entire completions based solely on final correctness, allowing unsupported evidence to be implicitly rewarded. Context-formatted RL [Yang et al., 2025, Chen et al., 2026] improves reasoning plausibility through structured chains-of-thought, but still leaves evidence unoptimized, where support remains implicit, untraceable, and cannot be directly attributed or improved. In both cases, the learning signal is entangled across the full response, making it unclear whether failures stem from incorrect evidence selection or flawed reasoning over valid evidence.

These observations suggest that the core bottleneck is not the lack of reasoning capacity, but the absence of a proper optimization target, where support is not treated as a first-class object in RL. While prior work exposes errors in temporal localization, modality attribution, and hallucinated evidence [Kim et al., 2024, Gong et al., 2024, Jung et al., 2025], and structured representations make evidence more explicit [Pu et al., 2026, Tencent Hunyuan Team, 2026], existing methods still do not optimize whether the model selects the correct, minimal, and sufficient evidence for each question.

To address this limitation, we reformulate AVQA as a support-centric optimization problem, where reasoning is decomposed into a sequence of support decisions and learning is driven by how well these decisions are made. We introduce OmniSE, a unified framework for support-faithful reinforcement learning in AVQA. Each QA instance is first transformed into an Audio-Visual Evidence Trace (AVET), which explicitly represents temporal evidence, modality-specific cues, and cross-modal relations. OmniSE optimizes this structure through three key components. First, we propose Support-Tree Rollout (STR) within a structured response space to separate support exploration from reasoning by sampling alternative evidence branches and generating reasoning-answer continuations conditioned on each branch. Second, factorized reward assignment is introduced to decompose the learning signal into multiple facets, including temporal localization, modality grounding, cross-modal binding, reasoning support and answer correctness, and then route each reward to the spans responsible for that decision. Third, a reward-gated self-consistency term is used to enforce that, once reliable support is established, downstream reasoning and answers remain stable across continuations. By elevating support to the primary unit of optimization, OmniSE enables models to produce not only correct answers, but answers grounded in faithful, minimal, and verifiable audio-visual evidence.

Our contributions are threefold:

- We introduce AVET, a structured representation of temporal, modality-specific, and cross-modal evidence to formulate AVQA as a support-faithful evidence-centric optimization problem.

- We propose OmniSE, a unified framework with Support-Tree Rollout, factorized reward routing, and reward-gated reasoning self-consistency regularization.
- Extensive experiments demonstrate that OmniSE consistently outperforms strong open-source baselines at comparable model scales.

2 Related Work

2.1 Omni-Modal LLMs

Omni-modal large language models extend multimodal generation beyond image-text settings by jointly processing video, audio, and text with synchronized temporal representations. Qwen2.5-Omni is a representative audio-video-capable backbone that supports end-to-end perception and generation across modalities [Xu et al., 2025a]. Recent structured audio-video interfaces further make long-form AV content more explicit: OmniScript produces temporally grounded hierarchical scripts, while Script-a-Video replaces monolithic captions with factorized and relationally grounded scene descriptions [Pu et al., 2026, Tencent Hunyuan Team, 2026]. In parallel, support-sensitive AVQA benchmarks and grounding studies show that model failures are often not simple recognition errors, but errors in temporal support selection, cross-modal attribution, and hallucinated evidence [Kim et al., 2024, Gong et al., 2024, Jung et al., 2025, Di and Xie, 2024, Zhang et al., 2023]. These lines of work provide the backbone, interface, and evaluation motivation for audio-video reasoning. OmniSE targets a different layer: it asks how the question-conditioned event, visual cue, audio cue, and audio-visual relation should be represented as the training object for QA.

2.2 RL for Omni-Modal LLMs

Reinforcement learning has become a major route for improving reasoning after supervised pre-training. In language models, GRPO and DAPO show that post-training can substantially improve reasoning behavior when the reward is well aligned with the task objective [Shao et al., 2024, Yu et al., 2025]. Recent omni-modal work transfers this idea to video and audio-video settings. Video-R1 and VideoChat-R1 adapt GRPO-style training to strengthen video reasoning and spatio-temporal perception, while EchoInk-R1, HumanOmniV2, and OmniVideo-R1 extend RL to open-world audio-video reasoning with stronger cross-modal or context-aware rewards [Feng et al., 2025, Li et al., 2025, Xing et al., 2025, Yang et al., 2025, Chen et al., 2026]. Related structured-reasoning and verifier-based work further shows the value of exposing or checking intermediate steps instead of judging only final answers [Zhang et al., 2024, Sun et al., 2025, Lightman et al., 2024]. These studies show that post-training and process signals can improve multimodal reasoning, but the reward and credit assignment are still usually attached to the completion, the answer, or a coarse process trace. OmniSE instead factorizes reward by AV support function and routes each component back to the response span that produced the corresponding localization, cue selection, relation binding, reasoning, or answer decision.

3 Method

OmniSE makes question-conditioned support the optimization interface for evidence-centric audio-video question answering (AVQA). For each audio-video QA instance, we build an Audio-Visual Evidence Trace (AVET) that decomposes answer support into temporal spans, modality-specific cues, audio-visual relations, and cited reasoning steps. After SFT on AVET data, the RL stage optimizes this support object with Support-Tree Rollout (STR): competing evidence branches are explored before reasoning continuations are compared within each branch. Support-factorized rewards separately score localization, modality evidence, audio-visual binding, reasoning support, and answer correctness, and route each signal to the response span that expresses the corresponding decision.

3.1 Audio-Visual Evidence Trace (AVET) Data Construction

Raw audio-video QA annotates final answers, but rarely specifies the cues, modalities, and cross-modal relations that justify them. We construct AVET targets with a lightweight two-stage annotation

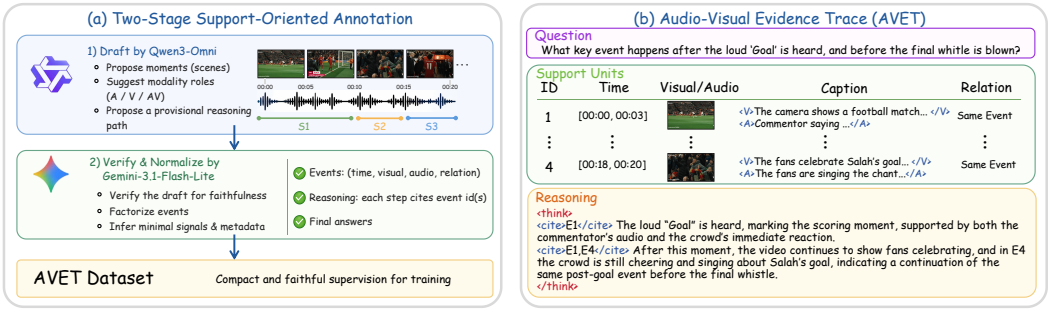


Figure 2: AVET data construction overview. Each QA sample is rewritten into an Audio-Visual Evidence Trace (AVET) with temporal evidence units, visual cues, audio cues, cross-modal relations and cited reasoning steps, making the minimal question-conditioned support explicit before SFT or reinforcement learning.

pass: Qwen3-Omni [Xu et al., 2025b] proposes candidate temporal moments, modality cues, audio-visual relations, and cited reasoning, and Gemini-3.1-Flash-Lite verifies the proposal and normalizes it into a compact support trace. The resulting target identifies the minimal support needed to answer: temporal evidence units, modality-specific cues, their audio-visual relation and cited reasoning steps. It preserves the support decisions used by SFT, STR, and support-factorized rewards without exposing auxiliary annotation diagnostics as model-facing targets.

After normalization, an AVET target serves as the common unit for supervision and reward computation. It pairs the answer $a^\star$ with a small set of support units $\mathcal{E}^\star$; each unit localizes the relevant interval and separates the visual cue, audio cue, and their relation. The cited reasoning trace $\mathcal{Z}^\star$ connects the answer to these units.

3.2 AVET-SFT Cold Start

AVET-SFT establishes a structured response protocol prior to reinforcement learning. It maintains a compact, evidence-conditioned reasoning interface while enriching the evidence field with temporal anchors, modality-specific cues, relation labels, and cited evidence-unit identifiers. The AVET-SFT explicitly encodes support structure. As a result, reinforcement learning focuses on refining support optimization, rather than inferring the schema from sparse answer-level feedback.

3.3 From GRPO to Audio-Visual Credit Assignment

GRPO is a convenient objective for reasoning post-training because it compares multiple completions sampled for the same prompt without requiring a separate value model [Shao et al., 2024, Yu et al., 2025]. For an input x, let $\hat{y}^{(1)}, \ldots, \hat{y}^{(G)}$ denote G sampled completions. Standard GRPO assigns each completion a scalar reward $r^{(g)}$, centers it within the group as

$$A^{(g)} = r^{(g)} - \frac{1}{G} \sum_{g'} r^{(g')},$$

and applies the same advantage to all tokens in $\hat{y}^{(g)}$. However, this completion-level credit assignment is too coarse for AVQA. A response may obtain the correct answer while grounding the wrong temporal span, relying on only one modality, or mentioning both modalities without establishing their support relation. In such cases, a single completion-level reward cannot tell whether failure comes from selecting the wrong support or from reasoning incorrectly over an otherwise plausible support. This ambiguity is especially problematic in audio-visual reasoning, where successful answering depends on multiple coupled decisions. A suitable objective therefore requires finer-grained credit assignment than a single scalar reward over the whole completion.

Our method is built around this observation. Rather than treating the entire completion as a single unit, STR structures the rollout around support selection and reasoning over the support units. Meanwhile, the factorized rewards assign credit to the specific response segments responsible for support grounding and subsequent reasoning.

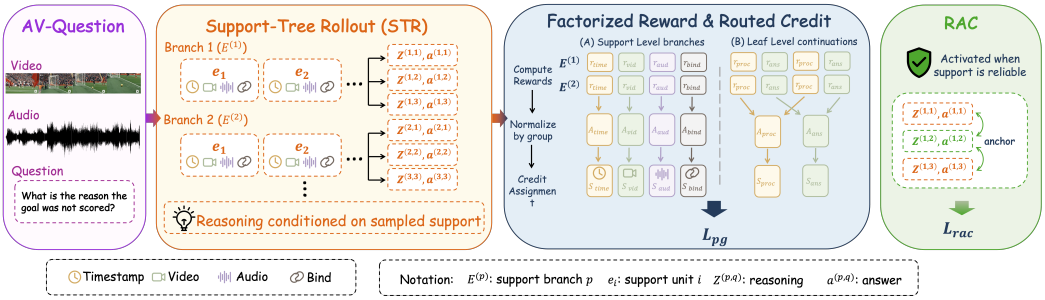


Figure 3: Overview of OmniSE. OmniSE first samples alternative audio-video support branches with Support-Tree Rollout, then generates support-conditioned reasoning and answers. Factorized rewards various facets, while routed credit assignment updates the corresponding response spans.

3.4 OmniSE: Support-Faithful Optimization

An overview of OmniSE is shown in Fig. 3. OmniSE preserves the group-relative comparison principle of GRPO [Shao et al., 2024], but replaces flat completion-level rollouts with a support-first factorization. Concretely, it consists of two coupled components: a support-tree rollout (STR), which first samples candidate support branches and then samples cited reasoning and answers conditioned on each branch; and a factorized reward assignment scheme, which evaluates support perception and support-conditioned reasoning at their respective levels and routes credit back to the response segments responsible for each decision.

3.4.1 Support-Tree Rollout

Structured response space. The OmniSE framework operates over an AVET response space that makes support decisions explicit. Each trace pairs the answer $a^\star$ with support units $\mathcal{E}^\star$, cited reasoning $\mathcal{Z}^\star$ and a modality label $s_{\text{req}} \in \{A, V, AV\}$. This response space defines the support branches, reasoning continuations, and token spans that STR later samples.

Support-tree rollout. GRPO compares completions sampled for the same prompt [Shao et al., 2024, Yu et al., 2025], but evidence-centric AVQA requires separating two decisions that a flat completion reward entangles: selecting valid support and reasoning from that support to the answer. STR therefore uses support as an explicit structural anchor for inference. For each prompt x, the policy first samples B support branches

$$\hat{\mathcal{E}}^{(p)} = \{\hat{e}_k^{(p)}\}_{k=1}^{K_p}, \quad \hat{e}_k^{(p)} = (\tau_k^{(p)}, v_k^{(p)}, u_k^{(p)}, \rho_k^{(p)}), \tag{1}$$

where $\hat{e}_k^{(p)}$ denote a support evidence unit that contain a temporal span, a visual cue, an audio cue, and a relation label. Conditioned on each support branch p, the policy then samples M continuations $(\hat{\mathcal{Z}}^{(p,q)}, \hat{a}^{(p,q)})$, where $\hat{\mathcal{Z}}^{(p,q)}$ denotes cited reasoning steps and $\hat{a}^{(p,q)}$ the final answer. Each cited reasoning step is represented as $\hat{z}_m^{(p,q)} = (\Gamma_m^{(p,q)}, h_m^{(p,q)})$, with citation set $\Gamma_m^{(p,q)}$ and step text $h_m^{(p,q)}$. This rollout yields $B \times M$ leaf responses,$\hat{y}^{(p,q)} = [\hat{\mathcal{E}}^{(p)}, \hat{\mathcal{Z}}^{(p,q)}, \hat{a}^{(p,q)}]$, but the comparison structure remains hierarchical rather than flat: support rewards compare branches across p, whereas process and answer rewards compare sibling continuations across q under the same support branch. Parsing each leaf produces token subsets $\mathcal{S}_{\text{time}}$, $\mathcal{S}_{\text{vis}}$, $\mathcal{S}_{\text{aud}}$, $\mathcal{S}_{\text{rel}}$, $\mathcal{S}_{\text{proc}}$, and $\mathcal{S}_{\text{ans}}$, which specify where the corresponding reward facets are applied. Missing or malformed fields yield empty subsets and receive zero credit.

3.4.2 Factorized Reward Assignment

Support rewards. Factorized reward assignment scores the support object before scoring the final answer. Given the reference evidence units $\mathcal{E}^\star = \{(\tau_j^\star, v_j^\star, u_j^\star, \rho_j^\star)\}_{j=1}^{K^\star}$ and a predicted support branch $\hat{\mathcal{E}}^{(p)}$, we assume both sequences are temporally ordered and perform one-to-one sequential

alignment between reference and predicted units. We then score the four support facets separately:

$$s_{j,p}^{\text{time}} = \text{IoU}(\tau_j^{(p)}, \tau_j^\star), \;\; s_{j,p}^{\text{vis}} = \text{Sim}_\phi(v_j^{(p)}, v_j^\star), \;\; s_{j,p}^{\text{aud}} = \text{Sim}_\phi(u_j^{(p)}, u_j^\star), \;\; s_{j,p}^{\text{rel}} = \mathbf{1}[\rho_j^{(p)} = \rho_j^\star]. \tag{2}$$

Here IoU is temporal intersection-over-union and $\text{Sim}_\phi(\cdot,\cdot) \in [0,1]$ is a pairwise semantic similarity score returned by a lightweight verifier LLM on short cue texts. The branch-level support rewards are simple averages over the matched reference units:

$$r_{\text{time}}^{(p)} = \frac{1}{K^\star}\sum_j s_{j,p}^{\text{time}}, \quad r_{\text{vis}}^{(p)} = \frac{1}{K^\star}\sum_j s_{j,p}^{\text{vis}}, \quad r_{\text{aud}}^{(p)} = \frac{1}{K^\star}\sum_j s_{j,p}^{\text{aud}}, \quad r_{\text{bind}}^{(p)} = \frac{1}{K^\star}\sum_j s_{j,p}^{\text{rel}}. \tag{3}$$

Reasoning and answer rewards. Reasoning is evaluated only through the support it cites. For each leaf (p,q), a lightweight process judge reads the cited evidence units, the step text, and the final answer, and returns normalized scores in $[0,1]$ for citation validity q_{cite}, local support q_{local}, unsupported new content q_{nonew}, and chain consistency q_{chain}:

$$r_{\text{proc}}^{(p,q)} = q_{\text{cite}}^{(p,q)} + q_{\text{local}}^{(p,q)} + q_{\text{nonew}}^{(p,q)} + q_{\text{chain}}^{(p,q)}. \tag{4}$$

The judge is restricted to short local contexts around cited support units, so it checks whether the reasoning is licensed by the selected support rather than acting as a general answer verifier. The answer reward is computed as:

$$r_{\text{ans}}^{(p,q)} = \text{Ans}\left(\hat{a}^{(p,q)}, a^\star\right), \tag{5}$$

where $\text{Ans}(\cdot,\cdot)$ is exact matching for multiple-choice questions and semantic similarity for open-ended questions.

Groupwise utility and routed credit assignment. GRPO assigns a single scalar reward to each completion and normalizes all completions within the same group. In contrast, we decompose each response into multiple reward facets and normalize them independently.

Each response contains multiple reward facets derived from the above support rewards, reasoning rewards and answer reward. Instead of mixing them into a single scalar, each facet forms its own comparable group and is normalized separately. Support facets compare branches, while reasoning and answer facets compare sibling continuations under the same support branch. For support facet $s \in \{\text{time}, \text{vis}, \text{aud}, \text{bind}\}$, we compute

$$u_s^{(p)} = \frac{1}{B-1}\sum_{p' \neq p}\left(2\sigma\left(\frac{\alpha_s(r_s^{(p)} - r_s^{(p')})}{T_s}\right) - 1\right), \quad A_s^{(p)} = u_s^{(p)} - \frac{1}{B}\sum_{p'=1}^{B} u_s^{(p')}. \tag{6}$$

For leaf facet $t \in \{\text{proc}, \text{ans}\}$, we center the utility among sibling continuations similarly, obtaining $u_t^{(p,q)}$ and $A_t^{(p,q)}$. The centered advantages are routed to tokens via facet-specific spans, except that the answer advantage is kept as a leaf-level completion signal. Let $\mathcal{S}_s$ denote the token set associated with support or process facet s. The token-level advantage is defined as

$$\tilde{A}_i^{(p,q)} = A_{\text{ans}}^{(p,q)} + \sum_{s \in \{\text{time},\text{vis},\text{aud},\text{bind}\}} \mathbf{1}[i \in \mathcal{S}_s] A_s^{(p)} + \mathbf{1}[i \in \mathcal{S}_{\text{proc}}] A_{\text{proc}}^{(p,q)}. \tag{7}$$

Here $A_{\text{ans}}^{(p,q)}$ provides a support-gated completion-level signal, while the non-answer facets are applied only to their corresponding token spans. The compact policy objective is

$$\mathcal{L}_{\text{pg}}(\theta) = -\mathbb{E}\left[\sum_{p=1}^{B}\sum_{q=1}^{M}\sum_i \min\left(r_i^{(p,q)}(\theta)\tilde{A}_i^{(p,q)}, \text{clip}\left(r_i^{(p,q)}(\theta), 1-\epsilon, 1+\epsilon\right)\tilde{A}_i^{(p,q)}\right)\right],$$
$$r_i^{(p,q)}(\theta) = \frac{\pi_\theta(\hat{y}_i^{(p,q)} \mid x, \hat{y}_{<i}^{(p,q)})}{\pi_{\theta_{\text{old}}}(\hat{y}_i^{(p,q)} \mid x, \hat{y}_{<i}^{(p,q)})}. \tag{8}$$

This design keeps answer correctness as a global leaf signal while still enabling fine-grained credit assignment: localization tokens are updated from time rewards, modality tokens from visual and audio rewards, binding tokens from relation rewards, and reasoning tokens from process rewards.

Method	Daily-Omni	WorldSense	IntentBench	VideoHolmes
Large-scale Models (≥30B / Closed-source)				
Gemini-2.0-Flash	67.8	56.2	67.8	30.6
Gemini-2.5-Pro	81.4	64.6	67.2	64.9
Gemini-3-Pro	81.1	66.4	71.5	67.0
Qwen3-Omni-30B-A3B-Instruct	63.6	54.0	65.7	49.0
Qwen3-Omni-30B-A3B-Thinking	75.8	48.0	68.5	57.3
OmniVideo-R1	82.8	65.8	74.2	62.9
7B-scale Open-source Models				
VideoLLaMA2-7B	35.2	25.4	–	35.2
Qwen2.5-Omni-7B	47.5	45.4	64.2	16.4
MiniCPM-o-7B	53.1	–	54.5	–
VITA-1.5-7B	–	36.9	54.2	–
Ola-7B	49.9	–	57.4	18.0
HumanOmniV2-7B	58.5	47.1	**69.3**	39.0
Omni-R1	46.8	44.1	63.5	30.0
AV-Reasoner	53.8	44.6	59.6	39.6
AVATAR	55.7	46.0	61.9	**45.1**
Ours (7B-scale)				
SFT baseline	55.6	47.8	65.1	23.2
Plain GRPO	55.1	46.5	62.8	23.4
OmniSE	**65.0** (+9.4)	**52.6** (+4.8)	67.5 (+2.4)	38.2 (+15.0)

(a) Audio-Video QA Benchmarks

Model	VMMMU	VERO	LVBench
Gemini1.5Pro	53.9	-	33.1
Qwen3-VL	68.7	39.6	58.0
Qwen3-Omni	64.1	38.1	50.2
AVATAR	53.7	36.6	38.4
OmniSE	56.8	37.8	42.2

(b) Video Benchmarks

Model	MMAU	MMSU	CV15↓
Gemini2.5 Pro	77.4	77.7	9.9
Qwen2-Audio	57.4	53.3	8.6
Qwen2.5-Omni	65.5	60.6	7.6
Qwen3-Omni	77.5	69.0	6.1
Ola	67.1	62.6	13.1
OmniSE	69.5	68.4	6.5

(c) Audio Benchmarks

Table 1: **(a)** Main comparison across audio-video QA benchmarks. Models are grouped by scale for fair comparison. Our method is based on a 7B model and is primarily compared with other 7B-scale systems. Blue numbers indicate absolute improvements over the SFT baseline. **(b)** Comparison across video-centric benchmarks. **(c)** Comparison across audio-centric benchmarks.

Reward-gated anchor consistency. The objective above supervises each sampled leaf independently, but does not enforce that sibling continuations under the same support branch produce stable inference outcomes. Under STR, all siblings share the same evidence hypothesis $\hat{\mathcal{E}}^{(p)}$. When this hypothesis is reliable, both the reasoning trace and the final answer should be largely determined by it, i.e., fixed support should imply stable inference. We therefore introduce a simple self-supervised regularization that reduces within-branch variation over reasoning and answer representations on reliable branches. We define a gated indicator that requires both reliable support and at least one high-quality continuation:

$$w^{(p)} = \mathbf{1}\left[r_{\text{sup}}^{(p)} > \delta_s \ \wedge \ \max_q \left(r_{\text{proc}}^{(p,q)} + \beta r_{\text{ans}}^{(p,q)}\right) > \delta_c\right], \tag{9}$$

where $r_{\text{sup}}^{(p)}$ is computed from the required support facets according to s_{req}. Within each active branch, we select the anchor continuation $q_p^\star = \arg\max_q \left(r_{\text{proc}}^{(p,q)} + \beta r_{\text{ans}}^{(p,q)}\right)$ and align all other siblings to its continuation representation. Let $h^{(p,q)} = \psi\left(\hat{y}^{(p,q)}_{\mathcal{S}_{\text{proc}} \cup \mathcal{S}_{\text{ans}}}\right)$, where $\psi(\cdot)$ denotes the encoding process of Sentence-BERT [Reimers and Gurevych, 2019], and $D(\cdot,\cdot)$ denotes cosine distance between textual embeddings. We minimize

$$\mathcal{L}_{\text{rac}} = \frac{1}{B}\sum_{p=1}^{B} w^{(p)} \cdot \frac{1}{M-1}\sum_{q \neq q_p^\star} D\left(h^{(p,q)}, \text{sg}(h^{(p,q_p^\star)})\right). \tag{10}$$

The final objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pg}} + \lambda_{\text{rac}}\mathcal{L}_{\text{rac}}. \tag{11}$$

4 Experiments

4.1 Experimental Setup

Training setup. To obtain high-quality training data, we sample part of the data from Video-R2 [Maaz et al., 2025], HumanOmniv2 [Yang et al., 2025] and VideoVista [Li et al., 2024]. We collect 26k high-quality AVET data for SFT cold-start. For the data used in the reinforcement learning stage, the data are sampled from the SFT data, according to the difficulty calculated by the answer

correctness from the rollouts of the SFT cold-start checkpoint. Qwen2.5-Omni [Xu et al., 2025a] is used as the backbone for experiments. Video sampling is fixed at 2 FPS, we use a maximum video frame number of 64 during SFT and 32 during RL. The learning rate is set to $1e^{-5}$ and $1e^{-6}$, respectively for the cold-starting training stage and RL stage, with a batch size of 8 and gradient accumulation steps of 4.

Evaluation setup. We evaluate on four audio-video QA benchmarks: Daily-Omni [Zhou et al., 2026], WorldSense [Hong et al., 2025], IntentBench [Yang et al., 2025], and VideoHolmes [Cheng et al., 2025]. We compare with a series of omni-modal RL models, including MiniCPM-o-7B [Yao et al., 2024], VITA-1.5-7B [Fu et al., 2025], Ola-7B [Liu et al., 2025], HumanOmniV2-7B [Yang et al., 2025], Omni-R1 [Zhong et al., 2025], AV-Reasoner [Lu et al., 2025], AVATAR [Kulkarni and Fazli, 2025] and OmniVideo-R1 [Chen et al., 2026]. We also evaluate OmniSE on the visual-centric and audio-centric benchmarks, including Video-MMMU [Hu et al.], Vript-ERO [Yang et al., 2024] and LVBench [Wang et al., 2025b] for video capacity, and MMAU [Sakshi et al., 2024], MMSU [Wang et al., 2025a] and CV15 [Ardila et al., 2020] for audio capacity. Evaluation is deterministic, with a temperature as zero and the same output-length budget used across compared methods. In testing, the video frame sampling rate is 2 FPS, with a maximum of 32 frames. More details about the evaluation benchmark can be found in the appendix.

4.2 Main Results Across Benchmarks

Table 1(a) reports the main results on audio-video benchmarks. Under the same Qwen2.5-Omni [Xu et al., 2025a] backbone, plain GRPO does not consistently improve over SFT (e.g., 55.1 vs. 55.6 on Daily-Omni and 46.5 vs. 47.8 on WorldSense), indicating limited benefit from naive answer-centric RL. In contrast, OmniSE yields consistent gains across all benchmarks, improving Daily-Omni from 55.6 to 65.0 (+9.4) and WorldSense from 47.8 to 52.6 (+4.8), while outperforming strong 7B baselines such as HumanOmniV2-7B [Yang et al., 2025] and AVATAR [Kulkarni and Fazli, 2025]. Despite using only a 7B backbone, OmniSE approaches or surpasses several 30B-scale models (e.g., 65.0 vs. 63.6 on Daily-Omni), suggesting that the gains arise from improved question-conditioned audio-video reasoning rather than scaling. As shown in Table 1(b),(c), OmniSE also achieves competitive performance on uni-modal benchmarks, indicating strong general perception and reasoning ability.

4.3 Ablation Study and Analysis

Ablation of each components. In this section, we conduct an ablation study to verify the effectiveness of each component within OmniSE in Tab. 2(a). As seen, removing the factorized reward results in the largest performance drop (e.g., -10.2 on Daily-Omni [Zhou et al., 2026] and -12.0 on VideoHolmes [Cheng et al., 2025]), indicating its central role in providing fine-grained supervision. Excluding routed credit assignment leads to moderate but consistent degradation, suggesting its importance in properly attributing learning signals across token spans. Finally, removing tree-level consistency regularization yields smaller yet noticeable declines, reflecting its role in stabilizing support-conditioned reasoning.

Effect of LLM-guided reasoning reward. We replace the LLM-as-judge reward with a lexical F1 score against captions of selected support units. As shown in Tab. 2(b), the F1-based variant still improves over the base model but consistently underperforms the LLM-guided reward, with a notable gap on VideoHolmes, which requires long-range compositional reasoning. This suggests that lexical overlap is a weak proxy for reasoning quality: models can achieve high F1 by copying captions without forming coherent inference chains. In contrast, the LLM-guided reward better captures whether the reasoning trace correctly grounds the final answer in the selected evidence.

Structured response space optimization improves the reasoning robustness. We evaluate the OmniSE model on two cross-modal hallucination benchmarks [Kim et al., 2024, Leng et al., 2024], to investigate the impact of the structured response space and tree-level optimization strategy on hallucination mitigation. As shown in Tab. 2(c), OmniSE consistently improves over the base model across AVHBench [Kim et al., 2024] and CMM [Leng et al., 2024], demonstrating that structuring the response space and performing tree-level optimization enhances the model's ability to produce faithful cross-modal reasoning, effectively reducing hallucinations.

Method	Daily-Omni	WorldSense	IntentBench	VideoHolmes
Qwen2.5-Omni	47.5	45.4	64.2	16.4
AVET-SFT	55.6	47.8	65.1	23.2
OmniSE	**65.0**	**52.6**	**67.5**	**38.2**
w/o factorized reward	54.8$_{(-10.2)}$	47.2$_{(-5.4)}$	65.4$_{(-2.1)}$	26.2$_{(-12.0)}$
w/o routed credit	62.5$_{(-2.5)}$	50.8$_{(-1.8)}$	65.8$_{(-1.7)}$	34.7$_{(-3.5)}$
w/o RAC	61.4$_{(-3.6)}$	51.6$_{(-1.0)}$	66.9$_{(-0.6)}$	37.6$_{(-0.6)}$

(a) Ablation study of method components.

Method	D-O	WS	IB	VH
F1-score	62.4	50.8	66.4	34.2
OmniSE	**65.0**	**52.6**	**67.5**	**38.2**

(b) Ablation on reasoning choices.

Model	AVHBench	CMM pa	CMM hr
Qwen2.5-Omni	72.16	91.74	75.28
OmniSE	**74.86**	**92.86**	**75.46**

(c) Hallucination Evaluation.

Table 2: **(a)** Component ablation across audio-video QA benchmarks. RAC denotes the reward-gated consistency regularization. **(b)** Ablation of reasoning rewards. D-O: Daily-Omni; WS: WorldSense; IB: IntentBench; VH: VideoHolmes. **(c)** Comparison of baseline model and OmniSE on two hallucination benchmarks: AVHBench [Kim et al., 2024] and CMM [Leng et al., 2024]. pa and hr denote perception accuracy and hallucination resistance, respectively.

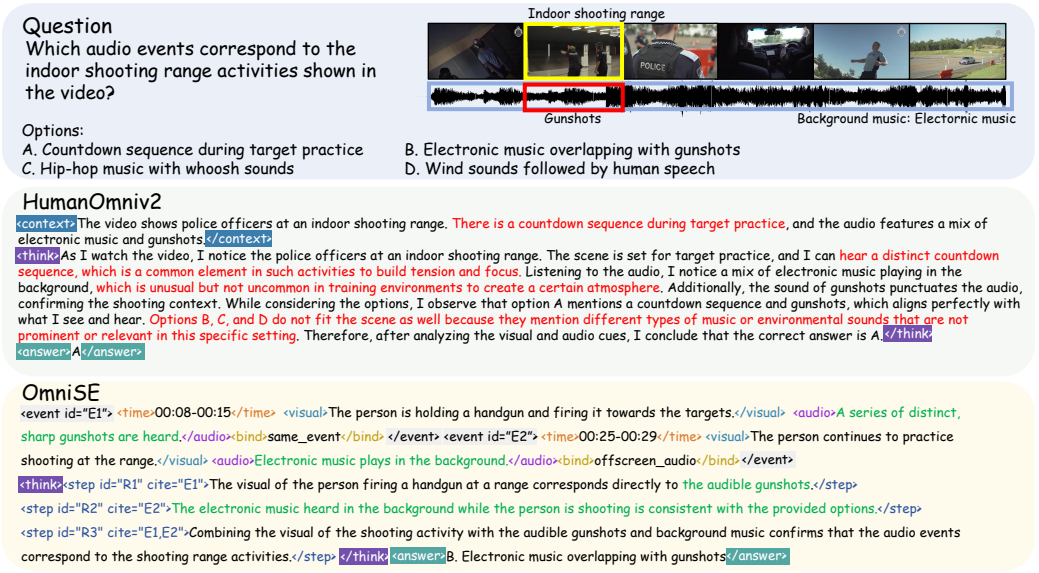


Figure 4: Qualitative visualization OmniSE's reasoning trace and comparison with context-formatted RL model HumanOmniV2 [Yang et al., 2025] with a sample from Daily-Omni [Zhou et al., 2026], demonstrating superior uni-modal perception and cross-modal reasoning ability from OmniSE. Red indicates hallucinated uni-modal perception and text prior reliance-driven reasoning trace. Green indicates accurate uni-modal grounding and effective cross-modal reasoning.

Qualitative case study. Fig. 4 illustrates the qualitative comparison of the context-driven RL method HumanOmniV2 [Yang et al., 2025] and our evidence-centric framework OmniSE. HumanOmniV2 produces a plausible but incorrect answer by relying on text priors and hallucinating a countdown sequence not supported by the audio. In contrast, with support unit being explicitly optimized in the RL process, OmniSE conducts its reasoning in temporally aligned visual and audio evidence, correctly identifying gunshots and background electronic music, and arrives at the correct answer through explicit cross-modal verification.

5 Conclusion

We presented OmniSE, a support-faithful optimization framework for evidence-centric audio-video question answering that reformulates AVQA as a support-centric problem. By representing question-conditioned evidence with audio-visual evidence traces and optimizing it via support-first rollout and factorized credit assignment, OmniSE separates support selection from support-conditioned reasoning and enables fine-grained supervision over temporal grounding, modality evidence, and

cross-modal relations. Experiments demonstrate that directly optimizing support leads to consistent gains over answer-level training objectives, indicating that correct multimodal reasoning requires not only accurate answers but also faithful and stable perception in the underlying evidence. These results suggest that support-aware representations and optimization provide a promising direction for improving evidence-centric reasoning in omni-modal models.

6 Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (No. 62306261), HK RGC-Early Career Scheme (No. 24211525) and ITSP Platform Project (No. ITS/600/24FP). This study was supported in part by the Centre for Perceptual and Interactive Intelligence, a CUHK-led InnoCentre under the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government. This work is also partially supported by Hong Kong RGC Strategic Topics Grant (No. STG1/E-403/24-N and No. STG1/M-404/26-N), and CUHK-CUHK(SZ)-GDST Joint Collaboration Fund (No. YSP26-4760949).

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 *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222, 2020.

Zhangquan Chen, Jiale Tao, Ruihuang Li, Yihao Hu, Ruitao Chen, Zhantao Yang, Xinlei Yu, Haodong Jing, Manyuan Zhang, Shuai Shao, Biao Wang, Qinglin Lu, and Ruqi Huang. Omnivideo-r1: Reinforcing audio-visual reasoning with query intention and modality attention. *arXiv preprint arXiv:2602.05847*, 2026. URL `https://arxiv.org/abs/2602.05847`.

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.

Shangzhe Di and Weidi Xie. Grounded question-answering in long egocentric videos. *arXiv preprint arXiv:2312.06505*, 2024. URL `https://arxiv.org/abs/2312.06505`.

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. *arXiv preprint arXiv:2503.21776*, 2025. URL `https://arxiv.org/abs/2503.21776`.

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*, 2025.

Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, and Xiangyu Yue. Av-odyssey bench: Can your multimodal llms really understand audio-visual information? *arXiv preprint arXiv:2412.02611*, 2024. URL `https://arxiv.org/abs/2412.02611`.

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.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos.(2025). *URL https://arxiv. org/abs/2501.13826*, 3.

Chaeyoung Jung, Youngjoon Jang, and Joon Son Chung. Avcd: Mitigating hallucinations in audio-visual large language models through contrastive decoding. *arXiv preprint arXiv:2505.20862*, 2025. URL `https://arxiv.org/abs/2505.20862`.

Sung-Bin Kim, Hyun-Bin Oh, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. *arXiv preprint arXiv:2410.18325*, 2024. URL `https://arxiv.org/abs/2410.18325`.

Yogesh Kulkarni and Pooyan Fazli. Avatar: Reinforcement learning to see, hear, and reason over video. *arXiv preprint arXiv:2508.03100*, 2025.

Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. *arXiv*, 2024. URL https://arxiv.org/abs/2410.12787.

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

Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. Videovista: A versatile benchmark for video understanding and reasoning. *arXiv preprint arXiv:2406.11303*, 2024.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

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*, 2025.

Lidong Lu, Guo Chen, Zhiqi Li, Yicheng Liu, and Tong Lu. Av-reasoner: Improving and benchmarking clue-grounded audio-visual counting for mllms. *arXiv preprint arXiv:2506.05328*, 2025.

Muhammad Maaz, Hanoona Rasheed, Fahad Shahbaz Khan, and Salman Khan. Video-r2: Reinforcing consistent and grounded reasoning in multimodal language models. *arXiv preprint arXiv:2511.23478*, 2025.

Junfu Pu, Yuxin Chen, Teng Wang, and Ying Shan. Omniscript: Towards audio-visual script generation for long-form cinematic video. *arXiv preprint arXiv:2604.11102*, 2026. URL https://arxiv.org/abs/2604.11102.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Linzhuang Sun, Hao Liang, Jingxuan Wei, Bihui Yu, Tianpeng Li, Fan Yang, Zenan Zhou, and Wentao Zhang. Mm-verify: Enhancing multimodal reasoning with chain-of-thought verification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14100–14115, 2025. doi: 10.18653/v1/2025.acl-long.689. URL https://aclanthology.org/2025.acl-long.689/.

Tencent Hunyuan Team. Script-a-video: Deep structured audio-visual captions via factorized streams and relational grounding. *arXiv preprint arXiv:2604.11244*, 2026. URL https://arxiv.org/abs/2604.11244.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. *arXiv preprint arXiv:2506.04779*, 2025a.

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

Zhenghao Xing, Xiaowei Hu, Chi-Wing Fu, Wenhai Wang, Jifeng Dai, and Pheng-Ann Heng. Echoink-r1: Exploring audio-visual reasoning in multimodal llms via reinforcement learning. *arXiv preprint arXiv:2505.04623*, 2025. URL https://arxiv.org/abs/2505.04623.

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. *arXiv preprint arXiv:2503.20215*, 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.

Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. *Advances in Neural Information Processing Systems*, 37:57240–57261, 2024.

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

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.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. URL https://arxiv.org/abs/2503.14476.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817, 2019.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Temporal sentence grounding in videos: A survey and future directions. *arXiv preprint arXiv:2201.08071*, 2023. URL https://arxiv.org/abs/2201.08071.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2024. URL https://arxiv.org/abs/2302.00923.

Hao Zhong, Muzhi Zhu, Zongze Du, Zheng Huang, Canyu Zhao, Mingyu Liu, Wen Wang, Hao Chen, and Chunhua Shen. Omni-r1: Reinforcement learning for omnimodal reasoning via two-system collaboration. *arXiv preprint arXiv:2505.20256*, 2025.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities, 2026. URL https://arxiv.org/abs/2505.17862.

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

Appendix of OmniSE: Support-Faithful Optimization for Evidence-Centric Audio-Video Reasoning

A More Implementation Details

A.1 Training Setup

We adopt the pretrained Qwen2.5-Omni [Xu et al., 2025a] as the backbone model for all experiments. The model is first initialized with supervised fine-tuning (SFT) on the constructed Audio-Visual Evidence Trace (AVET) dataset, followed by reinforcement learning using the proposed OmniSE framework.

For SFT, we train on approximately 26k AVET samples. The learning rate is set to 1×10^{-5}, with a batch size of 8 and gradient accumulation of 4 steps. We use the AdamW optimizer with beta set as 0.9 and 0.999, and epsilon as $1e-8$.

After SFT, we use the SFT checkpoint to do 8 rollouts for each SFT sample, and we calculate the answer correctness and the support reward of these rollouts to determine if a sample is well-learned during SFT. For the samples with an answer correctness lower than 75% and the support reward lower than the average of that across the whole SFT training set, they are denoted as hard samples and would be curated for RL training. However, due to computational resource constraints, we further randomly sample 3k data from these samples to form the formal RL training set. For the reinforcement learning stage, we initialize from the SFT checkpoint and apply Support-Tree Rollout (STR) with multiple sampled support branches and reasoning continuations per input. The learning rate is reduced to 1×10^{-6}. For the support-branch and continuation-reasoning branch, we set both the branch-level rollout numbers to 4 for these two numbers, leading to 16 rollouts per sample.

We set equal reward coefficients of 1 for each support facet. For reward-gated anchor consistency regularization, we set different answer reward thresholds for different types of questions. For a multiple-choice question, the question has to be correct, so the answer reward threshold is set to 1. For open-ended QA questions, where the reward is the F1 score with the annotation answer, the reward threshold is set as 0.4. For support reward gating, we set the support gating to 2. For the weight of the consistency regularization, we set it as 0.05. We set the max pixels for each frame as 100352.

All the experiments are run on the AMD MI210 GPUs.

A.2 Evaluation Setup

During evaluation, decoding is deterministic with the temperature set to 0. The maximum number of generated tokens is capped at 1024. Video inputs are sampled at 2 FPS, with a maximum of 32 frames per sample.

A.3 Reward Setup

For each sampled support branch $\hat{\mathcal{E}}^{(p)}$, we compute branch-level support rewards by aligning the predicted evidence units with the reference units $\mathcal{E}^\star = \{(\tau_j^\star, v_j^\star, u_j^\star, \rho_j^\star)\}_{j=1}^{K^\star}$. The temporal reward measures localization quality as $r_{\text{time}}^{(p)} = \frac{1}{K^\star}\sum_j \text{IoU}(\tau_j^{(p)}, \tau_j^\star)$. The visual and audio rewards measure whether the predicted modality-specific cues match the reference cues, computed as $r_{\text{vis}}^{(p)} = \frac{1}{K^\star}\sum_j \text{Sim}_\phi(v_j^{(p)}, v_j^\star)$ and $r_{\text{aud}}^{(p)} = \frac{1}{K^\star}\sum_j \text{Sim}_\phi(u_j^{(p)}, u_j^\star)$, where Sim_ϕ is a lightweight semantic similarity scorer over short cue descriptions. The binding reward evaluates cross-modal relation correctness as $r_{\text{bind}}^{(p)} = \frac{1}{K^\star}\sum_j \mathbf{1}[\rho_j^{(p)} = \rho_j^\star]$. The binding categories are AV, after, audio disambiguates visual, audio only, before, offscreen audio, parallel, same event, visual disambiguates audio, and visual only. We use the Qwen2.5-0.5B-Instruct as the semantic similarity scorer. The prompt for semantic similarity scores is shown in Fig. 5.

For each continuation (p, q) under this support branch, the process reward is produced by an LLM-as-judge over the cited evidence units and reasoning trace, $r_{\text{proc}}^{(p,q)} = q_{\text{cite}}^{(p,q)} + q_{\text{local}}^{(p,q)} + q_{\text{nonew}}^{(p,q)} + q_{\text{chain}}^{(p,q)}$, which checks citation validity, local support, absence of unsupported content, and chain consistency.

You are an English answer-grading judge for question answering. Read the English instructions carefully and output only one JSON object with one field: score. The score must be a number between 0 and 1. Use 1.0 only when the predicted answer matches the reference answer in meaning for the given question. Give a high score to correct paraphrases. Give a low score to contradictions, missing key facts, partially correct answers, or irrelevant answers. Be strict for numbers, counts, yes/no, dates, names, and option labels: they must match exactly.

Figure 5: Prompt for Qwen2.5-0.5B-Instruct for semantic similarity score.

For the process reward, we use the Qwen/Qwen3-VL-235B-A22B-Instruct-FP8 for LLM-as-judge. Prompt is shown in Fig. 6.

A.4 Data Setup

Training data is collected from VideoR2 [Maaz et al., 2025], Social-IQ [Yang et al., 2025] and VideoVista-Train [Li et al., 2024]. The total training data size is 26k, with 50% of them coming from VideoR2 [Maaz et al., 2025]. Notably, since we are working on the audio-video setting, with most of the training data collected from the video understanding works, we also prompt Gemini-3.1-Flash-lite to create audio-centric and cross-modal questions to enrich the variation of the training samples. 30% of the VideoVista-Train [Li et al., 2024] is randomly sampled from the high-quality clean subset from OmniVideo-R1 [Chen et al., 2026] and the remaining data from the Social-IQ [Zadeh et al., 2019] is randomly sampled to enrich the diversity of the training data. Prompts for Qwen3-Omni and Gemini-3.1-Flash-Lite as shown in Figs. 7, 8, respectively. Due to human resources limitations, the authors have only conducted manual verification on the data quality for 500 randomly sampled data points.

B Benchmark Details

We evaluate OmniSE on a diverse set of benchmarks covering audio-visual, visual-only, and audio-only reasoning tasks. These benchmarks are selected to comprehensively assess temporal grounding, cross-modal reasoning, and modality-specific understanding.

B.1 Audio-Visual Question Answering Benchmarks

Daily-Omni [Zhou et al., 2026]. Daily-Omni is an audio–visual question answering benchmark comprising 684 real-world videos collected from diverse sources. The dataset features rich multimodal (audio and visual) content and includes 1,197 multiple-choice question–answer pairs spanning six core task categories.

WorldSense [Hong et al., 2025]. WorldSense is a benchmark designed to evaluate omnimodal collaboration, featuring tightly coupled audio–visual tasks that necessitate synergistic multimodal reasoning. It comprises 1,662 synchronized audio–visual videos spanning eight domains and 67 subcategories, along with 3,172 multiple-choice question–answer pairs covering 26 tasks for comprehensive evaluation.

IntentBench [Yang et al., 2025]. IntentBench evaluates the capability of models to infer human intent and high-level semantics from audio-visual inputs. It requires reasoning over actions, goals, and interactions, often involving implicit cues that span both modalities. It comprises 633 videos and 2,689 questions grounded in both auditory and visual modalities.

VideoHolmes [Cheng et al., 2025] VideoHolmes is a challenging benchmark designed for compositional and multi-step reasoning in video understanding. It includes complex scenarios where answering questions requires long-range temporal reasoning and integrating multiple pieces of evidence across audio and visual streams. It contains 1837 questions from 270 annotated suspense short films across seven tasks.

B.2 Video-Centric Benchmarks

Video-MMMU [Hu et al.] Video-MMMU is a multimodal benchmark designed to evaluate knowledge acquisition and reasoning from professional, lecture-style videos across diverse disciplines. The benchmark consists of 300 expert-level educational videos spanning six domains, paired with 900

```
You are a visual-video and text-evidence judge for grounded audio-visual question answering.

Important modality limitation:
- You can inspect the visual stream of the video.
- You cannot hear, decode, or directly perceive audio.
- The model you are running on has no native audio perception ability.
- Therefore, all audio evidence must come only from provided text fields, such as ASR transcripts, audio captions, sound-event labels, or the "audio" fields
in support units.
- Never say or assume that you heard the audio.
- Never infer audio from visual content alone.

You will be given:
- A video input, used only for visual evidence.
- Textual audio evidence extracted by another system.
- The question and optional answer choices.
- Model-selected support units.
- The model reasoning trace.
- The model final answer.

Your task is to evaluate whether the reasoning is faithfully grounded in:
1. visible evidence from the video frames, and
2. textual audio evidence provided in the input.

Input:
- Visual video input: {video}
- Textual audio evidence:
{audio_text}
- Question: {question}
- Options, if any:
{options}
- Model-selected support units:
{support_units}
- Model reasoning trace:
{reasoning_trace}
- Model final answer:
{final_answer}

Evaluate four scores:

1. qcite: Citation validity
Check whether each reasoning step cites existing support unit IDs and whether the cited support units are relevant to the claim.
For visual claims, verify them against the video frames.
For audio claims, verify them only against the textual audio evidence.
For audio-visual claims, verify the visual side using video frames and the audio side using textual audio evidence.

2. qlocal: Local support
Check whether each reasoning claim is directly supported by the cited support units and available evidence.
For visual evidence, use the video frames around the cited timestamp.
For audio evidence, use only the textual audio evidence around the cited timestamp.
For cross-modal relations, verify that the visual evidence and textual audio evidence are temporally aligned or logically connected according to timestamps
and relation labels.

3. qnonew: No unsupported new content
Check whether the reasoning avoids introducing unsupported visual events, audio events, objects, speakers, temporal orderings, relations, intentions, or
causal claims.
Penalize any audio claim that is not explicitly present in the textual audio evidence.
Penalize any visual claim that is not visible in the video.

4. qchain: Chain consistency
Check whether the reasoning steps form a coherent and faithful path from the available evidence to the final answer.
The answer should follow from the cited visual evidence and textual audio evidence.

Judging rules:
- Judge reasoning faithfulness and grounding, not only final answer correctness.
- You may directly verify only visual claims from the video.
- You must verify audio claims only from textual audio evidence.
- Do not infer sounds from what is visually likely.
- Do not use outside knowledge or language priors.
- If a support unit contains an audio cue that is not supported by textual audio evidence, treat that audio cue as unreliable.
- If a support unit contains a visual cue that contradicts the video frames, treat that visual cue as unreliable.
- If a cross-modal relation depends on audio, verify the audio side only through textual audio evidence.
- A correct final answer does not compensate for ungrounded reasoning.
- Concise reasoning should not be penalized if it is fully supported.
- Be strict about timestamps, modality grounding, and option labels.
- Output valid JSON only.
- Do not include Markdown or explanations outside JSON.

Scoring scale:
- 1.0: fully valid, faithful, and grounded.
- 0.7: mostly grounded, with minor omissions or weak wording.
- 0.4: partially grounded, but with important unsupported or weak claims.
- 0.0: invalid, hallucinated, contradictory, or not assessable.

Return exactly:
{
  "qcite": <number between 0 and 1>,
  "qlocal": <number between 0 and 1>,
  "qnonew": <number between 0 and 1>,
  "qchain": <number between 0 and 1>
}
```

Figure 6: Prompt for Qwen/Qwen3-VL-235B-A22B-Instruct-FP8.

human-annotated question–answer pairs (three per video), covering three cognitive stages: perception, comprehension, and adaptation.

```
You are an audio-video evidence annotator for grounded audio-visual question answering.

Your task is to watch/listen to the given video and answer the question by constructing an Audio-Visual Evidence Trace (AVET). The AVET must explicitly identify the minimal temporal evidence, visual cues, audio cues, audio-visual relations, cited reasoning, and final answer.

You must not rely on language priors or guess unsupported events. Every reasoning step must be grounded in one or more evidence units.

Input:
- Video: {video}
- Question: {question}
- Options, if any: {options}
- Ground-truth answer, if available: {answer}

Annotation goal:
Produce a faithful AVET draft that contains only the evidence needed to answer the question.

Definitions:
1. support_unit:
   A temporally localized evidence unit that may contain:
   - visual evidence observed in the video
   - audio evidence heard in the video
   - the relation between the visual and audio evidence

2. relation_label:
   Choose exactly one label from:
   - same_event: audio and visual describe the same event at the same time
   - audio_disambiguates_visual: audio helps identify or interpret ambiguous visual content
   - visual_disambiguates_audio: visual helps identify or interpret ambiguous audio content
   - before_after: audio and visual evidence are temporally ordered and jointly answer the question
   - offscreen_audio: audio is relevant but its source is not visible
   - speaker_identity: visual speaker identity is linked to speech/audio
   - cross_scene_link: evidence must be linked across different temporal scenes
   - none: no meaningful audio-visual binding is required

3. modality_requirement:
   Choose exactly one from:
   - A: answer mainly requires audio evidence
   - V: answer mainly requires visual evidence
   - AV: answer requires both audio and visual evidence or their relation

Output requirements:
- Output valid JSON only.
- Do not include Markdown.
- Do not include explanations outside JSON.
- Use double quotes for all JSON keys and string values.
- Use timestamps in "HH:MM:SS.mmm-HH:MM:SS.mmm" format.
- Each support unit must have a unique id: "E1", "E2", ...
- Each reasoning step must cite existing support ids only.
- Do not invent objects, sounds, speakers, actions, or temporal events.
- If evidence is uncertain, write the uncertainty in the relevant cue field, but do not guess.
- Keep support units minimal and sufficient.
- Do not expose private diagnostic notes or chain-of-thought beyond concise cited reasoning steps.

Strict output schema:
{
  "question": "...",
  "options": {
    "A": "...",
    "B": "...",
    "C": "...",
    "D": "..."
  },
  "modality_requirement": "A|V|AV",
  "support_units": [
    {
      "id": "E1",
      "time": "HH:MM:SS.mmm-HH:MM:SS.mmm",
      "visual": "Concise visual cue observed in this interval. Use empty string if not relevant.",
      "audio": "Concise audio cue heard in this interval. Use empty string if not relevant.",
      "relation":
"same_event|audio_disambiguates_visual|visual_disambiguates_audio|before_after|offscreen_audio|speaker_identity|cross_scene_link|none"
    }
  ],
  "reasoning": [
    {
      "id": "R1",
      "cite": ["E1"],
      "text": "Concise support-grounded reasoning step. It must only use cited evidence."
    }
  ],
  "answer": {
    "label": "A|B|C|D|null",
    "text": "Final answer text"
  }
}

Quality checklist before output:
- Are all timestamps relevant to the question?
- Are support units minimal and sufficient?
- Does each reasoning step cite the evidence it depends on?
- Does the final answer follow from the cited evidence?
- Is there any hallucinated visual or audio content? If yes, remove it.
- Is the output valid JSON and nothing else?
```

Figure 7: Prompt for Qwen3-Omni for annotation initialization.

VERO (Vript-ERO) [Yang et al., 2024] VERO (Vript-ERO) is a video understanding benchmark designed to evaluate event recognition and temporal reasoning in real-world videos. It requires the model to do an event re-ordering task.

```
You are an audio-video evidence annotator for grounded audio-visual question answering.

Your task is to watch/listen to the given video and answer the question by constructing an Audio-Visual Evidence Trace (AVET). The AVET must explicitly identify the minimal temporal evidence, visual cues, audio cues, audio-visual relations, cited reasoning, and final answer.

You must not rely on language priors or guess unsupported events. Every reasoning step must be grounded in one or more evidence units.

Input:
- Video: {video}
- Question: {question}
- Options, if any: {options}
- Ground-truth answer, if available: {answer}

Annotation goal:
Produce a faithful AVET draft that contains only the evidence needed to answer the question.

Definitions:
1. support_unit:
   A temporally localized evidence unit that may contain:
   - visual evidence observed in the video
   - audio evidence heard in the video
   - the relation between the visual and audio evidence

2. relation_label:
   Choose exactly one label from:
   - same_event: audio and visual describe the same event at the same time
   - audio_disambiguates_visual: audio helps identify or interpret ambiguous visual content
   - visual_disambiguates_audio: visual helps identify or interpret ambiguous audio content
   - before_after: audio and visual evidence are temporally ordered and jointly answer the question
   - offscreen_audio: audio is relevant but its source is not visible
   - speaker_identity: visual speaker identity is linked to speech/audio
   - cross_scene_link: evidence must be linked across different temporal scenes
   - none: no meaningful audio-visual binding is required

3. modality_requirement:
   Choose exactly one from:
   - A: answer mainly requires audio evidence
   - V: answer mainly requires visual evidence
   - AV: answer requires both audio and visual evidence or their relation

Output requirements:
- Output valid JSON only.
- Do not include Markdown.
- Do not include explanations outside JSON.
- Use double quotes for all JSON keys and string values.
- Use timestamps in "HH:MM:SS.mmm-HH:MM:SS.mmm" format.
- Each support unit must have a unique id: "E1", "E2", ...
- Each reasoning step must cite existing support ids only.
- Do not invent objects, sounds, speakers, actions, or temporal events.
- If evidence is uncertain, write the uncertainty in the relevant cue field, but do not guess.
- Keep support units minimal and sufficient.
- Do not expose private diagnostic notes or chain-of-thought beyond concise cited reasoning steps.

Strict output schema:
{
  "question": "...",
  "options": {
    "A": "...",
    "B": "...",
    "C": "...",
    "D": "..."
  },
  "modality_requirement": "A|V|AV",
  "support_units": [
    {
      "id": "E1",
      "time": "HH:MM:SS.mmm-HH:MM:SS.mmm",
      "visual": "Concise visual cue observed in this interval. Use empty string if not relevant.",
      "audio": "Concise audio cue heard in this interval. Use empty string if not relevant.",
      "relation":
"same_event|audio_disambiguates_visual|visual_disambiguates_audio|before_after|offscreen_audio|speaker_identity|cross_scene_link|none"
    }
  ],
  "reasoning": [
    {
      "id": "R1",
      "cite": ["E1"],
      "text": "Concise support-grounded reasoning step. It must only use cited evidence."
    }
  ],
  "answer": {
    "label": "A|B|C|D|null",
    "text": "Final answer text"
  }
}

Quality checklist before output:
- Are all timestamps relevant to the question?
- Are support units minimal and sufficient?
- Does each reasoning step cite the evidence it depends on?
- Does the final answer follow from the cited evidence?
- Is there any hallucinated visual or audio content? If yes, remove it.
- Is the output valid JSON and nothing else?
```

Figure 8: Prompt for Gemini-3.1-Flash-Lite.

LVBench [Wang et al., 2025b] LVBench is a benchmark designed to evaluate long-video understanding, focusing on scenarios where models must process extended temporal contexts and perform

reasoning across long durations. The benchmark comprises 1,549 human-annotated question–answer pairs with an average video duration of 4,101 seconds, covering diverse domains such as TV shows, sports broadcasts, and real-world recordings.

B.3 Audio-Centric Benchmarks

MMAU [Sakshi et al., 2024] MMAU is a large-scale benchmark designed to evaluate multi-task audio understanding and reasoning across diverse acoustic scenarios. It covers a wide spectrum of audio-centric tasks, including acoustic event recognition, sound classification, audio captioning, and reasoning over auditory signals.

MMSU [Wang et al., 2025a] MMSU focuses on spoken language understanding and reasoning. It evaluates the ability to interpret speech content and perform reasoning over linguistic information in audio signals. The benchmark consists of 47 diverse tasks spanning multiple aspects of speech processing, including phonetic recognition, syntactic parsing, semantic understanding, and high-level reasoning over spoken content.

CV15 [Ardila et al., 2020] CV15 is a benchmark for audio perception and recognition, focusing on robustness and generalization across different acoustic conditions.

B.3.1 More Visualization

More visualizations of the model's responses are shown in Fig. 9.

C Limitations

C.1 Limitations of the Proposed Method

Although OmniSE provides a principled support-centric optimization framework for audio-video question answering, its method design still has several limitations.

Dependence on automatically constructed AVET supervision. OmniSE relies on Audio-Visual Evidence Trace (AVET) targets constructed through a two-stage annotation pipeline, where one model proposes temporal evidence, modality-specific cues, cross-modal relations, and reasoning traces, and another model verifies and normalizes them. This design reduces manual annotation cost, but it also makes the quality of training supervision dependent on the reliability of external annotator and verifier models. If the generated AVET contains incorrect temporal spans, incomplete modality cues, or biased relation labels, the subsequent SFT and RL stages may optimize toward noisy or misleading support structures.

Verifier and judge bias in reward computation. The factorized reward assignment uses lightweight semantic similarity scorers and LLM-as-judge modules to evaluate visual cues, audio cues, citation validity, local support, absence of unsupported content, and reasoning consistency. These judges may introduce their own biases, fail to detect subtle grounding errors, or reward plausible textual explanations that are not truly supported by the audio-video input. As a result, the learned policy may overfit to the preference patterns of the reward models rather than achieving genuine multimodal grounding.

Bottleneck for long videos. Since OmniSE implements a relatively complex output format, with a list of support units, when given long videos, the model might generate extremely long sequences of support units, leading to excessive token sequence length exceeding the model's context window and bringing large computational cost.

Lack of re-checking of the supprt units. OmniSE optimizes for the target of support-faithful optimization, which makes the assumption that the support generated previously is correct. However, this is not always the case in evaluation and real-world scenarios. Therefore, a potential re-checking and verification mechanism, with a potential paradigm of o3 [Zhu et al., 2025], could be helpful.

Question
Which event occurs FIRST in the video's timeline?



Options:
A. White fabric unrolled in a room with ambient music
B. Fabric being unrolled on a reflective surface with music
C. Liz Rose pushing a cart through a workshop doorway
D. 'Conservation' text overlay with blurred hands

OmniSE

```
<event id="E1">
<time>00:00-00:01</time>
<visual>A title card with the word 'Conservation' and blurred hands is displayed.</visual>
<audio><none></audio>
<bind>visual_only</bind>
</event>
<event id="E2">
<time>00:01-00:04</time>
<visual>White fabric is being unrolled on a reflective surface.</visual>
<audio>Ambient music plays.</audio>
<bind>parallel</bind>
</event>
<event id="E3">
<time>00:04-00:05</time>
<visual>Liz Rose pushing a cart through a doorway.</visual>
<audio><none></audio>
<bind>visual_only</bind>
</event>
<think>
<step id="R1" cite="E1">The video starts with the 'Conservation' title card, which is the first visual element.</step>
<step id="R2" cite="E2">The unrolling of fabric occurs after the title card.</step>
<step id="R3" cite="E3">Liz Rose pushing the cart happens later in the sequence.</step>
<step id="R4" cite="E1,E2,E3">Comparing the timestamps, the title card (E1) is the first event.</step>
</think>
<answer>D. 'Conservation' text overlay with blurred hands</answer>
```

Question
Why is the man urgently collecting the Agarikon fungus strain during this visit?



Options:
A. Optimal spore release occurs at this growth stage
B. The forest is scheduled for clear-cutting next year
C. It's a newly discovered species requiring documentation
D. The host log's rapid decay would spoil the specimen

OmniSE

```
<event id="E1">
<time>00:05-00:09</time>
<visual>The man is walking through a forest and mentions seeing a blue tag.</visual>
<audio>The man states that they saw a blue tag indicating the entire forest is going to be cut.</audio>
<bind>audio_disambiguates_visual</bind>
</event>
<event id="E2">
<time>00:22-00:26</time>
<visual>The man is standing in the forest, gesturing towards the area.</visual>
<audio>The man says, 'This area is going to be totally destroyed here in the next year. It's a perfect opportunity for us to get a another strain into culture.'</audio>
<bind>audio_disambiguates_visual</bind>
</event>
<think>
<step id="R1" cite="E1">The man identifies that the forest he is in is scheduled for clear-cutting within the next year, which creates a time-sensitive situation for collecting biological samples.</step>
<step id="R2" cite="E2">The man explicitly states that the urgency of collecting the Agarikon strain is because the area will be destroyed in the next year, making this the 'perfect opportunity' to collect it before it is too late.</step>
<step id="R3" cite="E1,E2">The combination of the forest's impending destruction and the man's stated goal of preserving the strain confirms that the forest's scheduled clear-cutting is the reason for the urgent collection.</step>
</think>
<answer>B. The forest is scheduled for clear-cutting next year</answer>
```

Figure 9: Visualization of OmniSE's responses.

D Societal Impacts

This work studies support-faithful optimization for grounded audio-video question answering, with the goal of encouraging omni-modal models to base their answers on temporally localized, modality-specific, and cross-modal evidence rather than on answer-only shortcuts or unsupported language priors. The potential positive societal impact of this direction is that it may improve the reliability and transparency of audio-video AI systems. By requiring models to identify the relevant time spans, visual cues, audio cues, and audio-visual relations that support an answer, the proposed framework can make model behavior more interpretable and easier to audit. This may be beneficial in applications

such as assistive video understanding, educational content analysis, accessibility tools for people with sensory impairments, and safer human-AI interaction in settings where both auditory and visual evidence matter. In particular, reducing unsupported cross-modal reasoning and hallucinated evidence can help users better understand when a model's answer is grounded in the input and when it may be unreliable.

At the same time, the work may also have negative societal impacts if deployed without appropriate safeguards. More capable audio-video reasoning systems could be used for privacy-invasive surveillance, automated monitoring, or profiling of individuals in recorded environments. Since the method improves temporal localization and cross-modal evidence attribution, it may make it easier to extract fine-grained behavioral information from videos, including speech, actions, interactions, and contextual cues. In addition, even support-faithful models may still make errors in ambiguous, noisy, culturally specific, or sensitive scenarios, and users may over-trust structured evidence traces as proof of correctness. The use of LLM-based judges and automatically constructed evidence traces may also introduce hidden biases from the verifier models or training data. Therefore, deployment should include privacy-preserving data practices, consent-aware use of audio-video data, careful evaluation across demographic and cultural contexts, human oversight in high-stakes applications, and clear communication that grounded reasoning traces are model-generated explanations rather than definitive evidence.

E Declaration of LLM usage

Large language models were used throughout the research process, including literature exploration, idea refinement, writing and editing, code development, experiment implementation, debugging, and result analysis. All LLM-assisted outputs were reviewed, verified, and revised by the authors, who remain fully responsible for the correctness, originality, and integrity of the work.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discuss the limitations in appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: No theoretical results included.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Details are provided in implementation details section of the main paper and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Data and code will be open-source after peer review.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Details are illustrated in the implementation details section of the main paper and the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Not related.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Shown in appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper, in every respect, is with the NeurIPS Code of Ethics

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Shown in appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Don't have such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Relevant resources are all well cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: Crucial assets will be made publicly available after peer review.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Not related.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Not related.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Shown in appendix

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.