SEEN, HEARD, OR ASSUMED? CROSS-MODAL HALLUCINATIONS IN AUDIO-VISUAL LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite the rapid progress of audio-visual large language models (AVLLMs), hallucination remains an important reliability challenge in joint audio-visual reasoning. Existing audio-visual hallucination benchmarks emphasize final-answer correctness or answer changes after modifying one modality, while it remains unclear whether these evaluation results reflect correct and stable judgments about what is heard and seen. To jointly assess final-answer correctness, modality-judgment correctness, and cross-modal stability, we introduce AV-TRACE, a benchmark with 7,344 test examples that combines explicit modality judgments with matched audio-visual comparisons for modality-specific hallucination detection and cross-modal stability testing. AV-TRACE reveals that correct final answers can mask modality-judgment failures, and that higher modality-level correctness can coexist with poorer cross-modal stability. Supervised fine-tuning improves answer and judgment accuracy, yet correct judgments can still flip when only the other modality changes. We therefore propose Factorial Quartet Fine-Tuning (FQFT), which augments GRPO with a joint answer-judgment reward and factorial regularization of judgment margins across four target-presence states. On Qwen2.5-Omni-7B, compared with supervised fine-tuning, FQFT increases the task joint accuracy of answers and modality judgments from 76.29% to 84.39%, reduces the hallucination rate from 20.49% to 10.32%, and reduces the flip rate from 4.08% to 1.95%.

1 INTRODUCTION

Recent audio-visual large language models (AVLLMs) extend large language models to jointly process audio, video, and text for multimodal perception and reasoning (Xu et al., 2025a;b). Hallucination has been widely observed in language and multimodal models, where model outputs may contain incorrect or fabricated content (Li et al., 2023a; Manakul et al., 2023; Li et al., 2023b; Guan et al., 2024). In audio-visual reasoning, hallucination can further take a cross-modal form, where information from one modality may unduly influence the model's interpretation of another (Sun et al., 2024; Sudheer et al., 2026; Zhang et al., 2025).

However, existing audio-visual hallucination benchmarks mainly assess final-answer correctness or changes in model outputs after one modality is modified (Kim et al., 2025; Ye et al., 2026; Wen et al., 2026; Dong et al., 2026). These protocols capture answer-level errors and sensitivity to input changes, but final-answer correctness alone does not show whether the model's modality judgments are correct or whether they remain stable when only the other modality changes. To expose failure modes that such evaluations can miss, we introduce **AV-TRACE**, which combines explicit modality judgments with matched cross-modal comparisons (Figure 1). AV-TRACE comprises 7,344 human-verified test examples across two complementary evaluations: modality-specific hallucination detection and cross-modal stability testing. It reveals that correct final answers can mask modality-judgment failures, and that higher modality-level correctness can coexist with poorer cross-modal stability.

Single-input correctness does not characterize how a modality judgment depends on the two modalities. A correct judgment can be consistent with different cross-modal behaviors, which are indistinguishable from a single input alone. Across related inputs, however, these behaviors can be separated: the model should distinguish audio clips with and without the queried sound, or video clips with and without the queried visual content, while preserving its judgment when the queried modality remains fixed and only the other modality changes. Reliable modality judgments therefore require selective sensitivity: a judgment should respond to relevant changes in the modality being queried

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

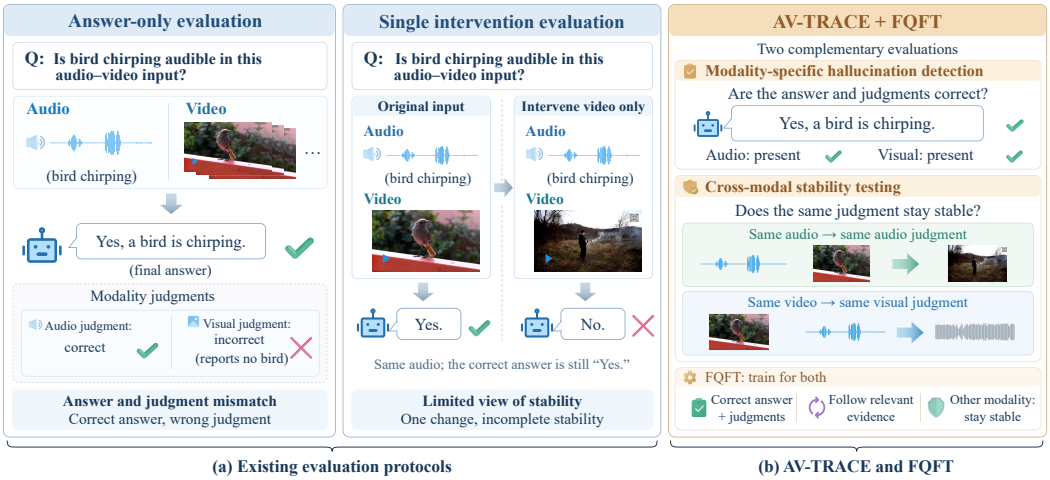


Figure 1: Motivation and overview of AV-TRACE and FQFT. (a) Answer-only evaluation can miss incorrect modality judgments despite a correct final answer, while a single intervention provides only a partial view of cross-modal stability. (b) AV-TRACE evaluates answer and modality-judgment correctness together with stability under matched cross-modal changes. FQFT trains judgments to follow evidence in the queried modality while remaining stable to changes in the other modality.

while remaining stable to changes in the other modality. Existing mitigation methods adjust modality contributions or use modified multimodal inputs to reduce inappropriate cross-modal reliance (Jung et al., 2025; Chung et al., 2026; Jung et al., 2026; Baid et al., 2026; Chaubey et al., 2026). For audio judgments on matched inputs, the same video change can shift judgment scores differently depending on whether the audio contains the queried sound. A comparison in either case combines the average video-associated variation with an interaction component that depends on the audio evidence. Relating the two comparisons separates these components. This motivates a joint objective that models modality-specific variation and interaction alongside judgment correctness.

To provide such an objective, we propose **Factorial Quartet Fine-Tuning (FQFT)**, which augments GRPO with a joint answer-judgment reward and a factorial quartet regularizer. The reward promotes correct answers and judgments, while the regularizer encourages target discrimination and score consistency across related audio-visual examples. On Qwen2.5-Omni-7B, compared with supervised fine-tuning, FQFT increases task-macro joint answer-judgment accuracy from 76.29% to 84.39%, reduces the cross-modal hallucination rate from 20.49% to 10.32%, and reduces the harmful flip rate from 4.08% to 1.95%. These results show that FQFT improves joint correctness and reduces the measured cross-modal hallucination and harmful flip rates.

Our contributions are:

- We introduce **AV-TRACE**, which pairs explicit modality judgments with matched cross-modal comparisons to assess modality-level correctness and cross-modal stability in a unified framework.
- AV-TRACE reveals a previously obscured separation between final-answer correctness, modality-level correctness, and cross-modal stability, showing that a correct judgment on a single input does not necessarily reveal how that judgment depends on the two modalities.
- We propose **Factorial Quartet Fine-Tuning (FQFT)**, which translates this distinction into a factorial objective for jointly learning modality judgments across related audio-visual examples.

2 Related Work

2.1 Audio-visual Hallucination Evaluation

Audio-visual hallucination evaluation has evolved from single-input diagnosis toward controlled comparisons that probe how model behavior changes across modalities. AVHBench evaluates hallucinations in both audio-to-visual and visual-to-audio directions, covering modality-specific failures under joint audio-visual input (Kim et al., 2025). Perturbation-based benchmarks extend this diagnosis by examining how responses change when the audio-visual input is modified. AV-ConfuseBench focuses on visually induced audio hallucinations by modifying sounds associated with

visible objects (Ye et al., 2026). Thud further studies sound existence, temporal synchronization, and audio-visual consistency through muting, shifting, and swapping operations (Wen et al., 2026). Rather than modifying the multimodal input, OmniHalluc-L pairs supported and counterfactual claims with fixed audio-visual evidence to evaluate evidence-binding hallucinations in long-form videos (Dong et al., 2026). Together, these benchmarks establish complementary single-input, perturbation-based, and counterfactual protocols for measuring cross-modal hallucination. Correctness and sensitivity nevertheless capture different aspects of reliability: a changed judgment can either introduce or correct an error. Relating the two therefore requires tracking whether a correct modality judgment remains correct when its audio or visual evidence is fixed and only the accompanying modality changes.

2.2 Audio-visual Hallucination Mitigation

Audio-visual hallucination mitigation has been studied at both inference and training time. Inference-time methods modify how information from different modalities contributes during generation without updating model parameters. AVCD contrasts predictions under attention-guided perturbations (Jung et al., 2025), MAD adaptively combines modality-conditioned decoding branches according to question relevance (Chung et al., 2026), and Adaptive Sink-Guided Decoding (ASD) rebalances attention between cross-modal and unimodal sink tokens for audio-visual captioning (Jung et al., 2026). Training-time approaches instead shape multimodal behavior through supervised alignment, preference optimization, reinforcement learning, or modality-specific reasoning. AVHBench studies supervised alignment and fine-tuning (Kim et al., 2025), while OmniDPO combines text and multimodal preference objectives (Chen et al., 2026). RL-CoMM uses audio-only teacher guidance together with intermediate reasoning rewards (Ye et al., 2026), whereas SFFL separates audio and visual reasoning before fusion (Li et al., 2026). More recent approaches incorporate modality changes into preference learning. ACPO combines preferences for audio-grounded over visually driven responses with comparisons of the same response under original and substituted audio (Baid et al., 2026). MoD-DPO promotes invariance to irrelevant-modality corruption and sensitivity to relevant-modality corruption through preference optimization (Chaubey et al., 2026). FQFT instead regularizes explicit judgment margins across four presence states, separating target-associated, other-label, interaction, and mean components alongside a joint answer-judgment reward.

3 Methods

AV-TRACE distinguishes the correctness of explicit modality judgments from their stability across matched audio-visual inputs. We first define its two complementary evaluations and scoring protocol, then present FQFT to jointly model modality-specific variation, interaction, and judgment correctness.

3.1 AV-TRACE

AV-TRACE contains 7,344 human-verified test examples across *modality-specific hallucination detection* and *cross-modal stability testing*. The first evaluation comprises 1,000 audio detection questions and 1,000 visual detection questions; the second comprises 5,344 inputs organized into 1,336 matched quartets. Every input contains both audio and video. Figure 3 summarizes the construction pipeline, and Table 1 compares the evaluation structure with related benchmarks.

Table 1: Audio-visual benchmarks. A/V: audio/visual; NR: unreported; –: no presence judgment. *CMM: complement of correct rejection. †Quartets. AVHBench counts omit captioning; rates use native definitions.

Benchmark	Questions	Response	Modality judgments	Evaluation structure	Hall. rate	Flip rate
AVHBench (2025)	5,302	Yes/no	A or V	Single input	NR	NR
CMM (2025)	2,400	Yes/no	A or V	Single input	✓*	NR
Thud (2026)	710	Open	Audio	Before/after audio change	✓	NR
OmniHalluc-L (2026)	3,600	Yes/no	–	Supported/counterfactual claim pair	NR	NR
AV-TRACE (ours)	**7,344**	4-way MCQ	**A and V†**	**Four matched inputs†**	✓	✓

Modality-specific hallucination detection. Each question asks which of four candidates is present in the queried audio or video, with one verified positive and three verified negatives. For *audio detection*, one designated negative names a sound semantically associated with the visible content but absent from the audio, providing a visual-to-audio distractor. For *visual detection*, this is reversed:

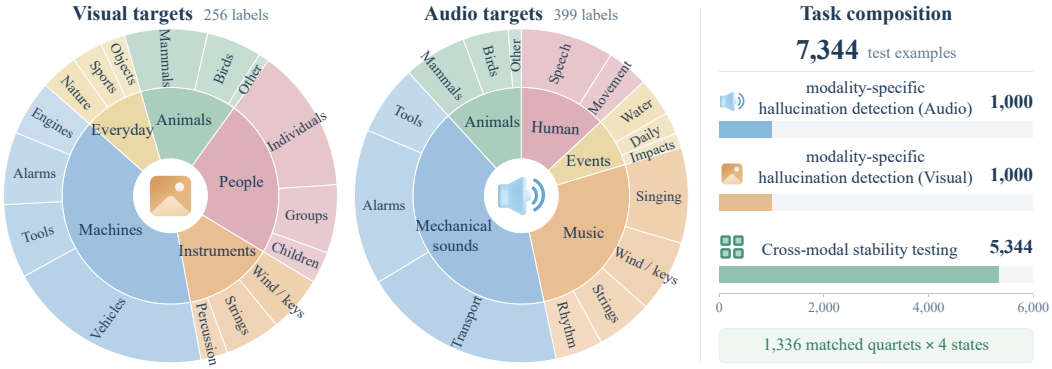


Figure 2: Semantic coverage and evaluation composition of AV-TRACE. Audio and visual categories span 399 and 256 labels, respectively. The evaluation includes 1,000 audio detection questions, 1,000 visual detection questions, and 1,336 matched quartets (5,344 inputs).

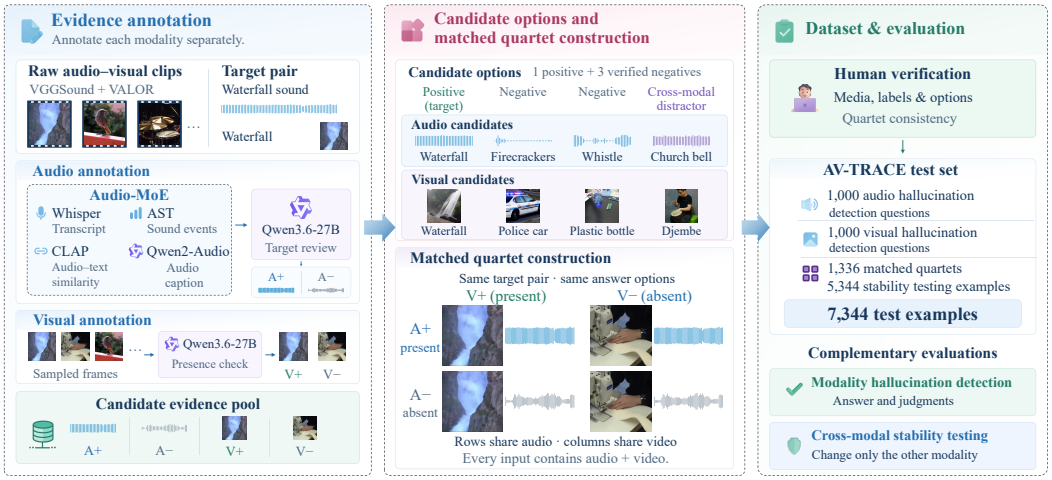


Figure 3: AV-TRACE construction pipeline. Targets are verified separately in audio and video, then assembled into detection questions with cross-modal distractors and matched quartets with shared clips. Human review checks the media, labels, options, and quartet consistency. The resulting 7,344 test examples support modality-specific hallucination detection and cross-modal stability testing; every input contains both modalities.

an audible cue suggests content absent from the video. The model judges each candidate in the queried modality before selecting a final answer. Candidate-to-option mappings are permuted across questions. Scoring candidate judgments separately from the final answer reveals whether a correct choice coexists with incorrect judgments about what is heard or seen. Figure 4 shows both directions.

Cross-modal stability testing. For the queried sound, we select one audio clip containing it and one without it. Likewise, we select one video clip showing the queried visual content and one without it. Cross-pairing the two audio clips with the two video clips produces a matched quartet: Both, Audio-only, Visual-only, and Neither. Every input includes both audio and video; the four labels specify whether the audio contains the queried sound and whether the video shows the queried visual content. Each input requests separate judgments of the queried sound and visual content, their combined state, and a final option encoding that state. The queried sound and visual content, question structure, and option mapping remain fixed within the quartet.

Each audio clip is paired with both videos, and each video with both audio clips. The resulting comparisons keep the queried-modality clip identical while changing only the other modality. This separates correctness on an individual input from stability across matched inputs. Figure 5 illustrates a harmful visual-to-audio flip; Section 3.2 uses the same quartet categories to formulate the training objective.

Construction and verification. AV-TRACE draws on VGGSound (Chen et al., 2020) and VALOR (Liu et al., 2025). Candidate sounds and visual content are verified separately, followed by two independent annotators and third-reviewer adjudication of disagreements. Training, validation,

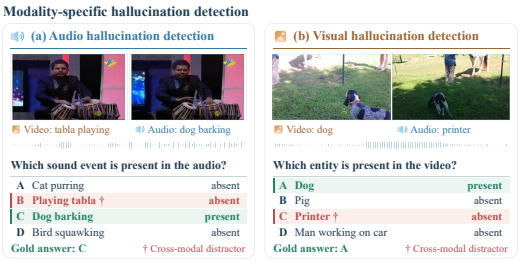


Figure 4: Modality-specific hallucination detection with cross-modal distractors. (a) The video shows tabla playing, but the audio contains dog barking (C), not tabla sounds (B). (b) The audio contains printer sounds, but the video shows a dog (A), not a printer (C). Each question provides both modalities and queries one; the presence labels describe the queried modality. Red marks cross-modal distractors, and green marks the correct candidates and answers.

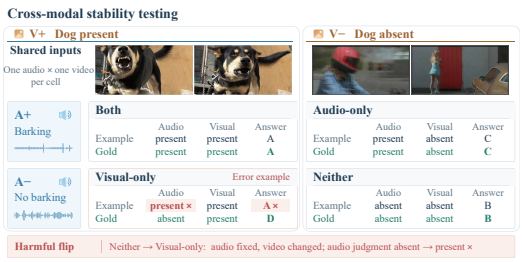


Figure 5: Cross-modal stability testing with a matched quartet. The targets are *dog barking* in audio and *dog* in video; rows reuse audio, columns reuse video, and all four inputs share the same answer mapping. In the Neither-to-Visual-only transition, only the video changes, but the illustrative audio judgment flips from absent to present and the final answer from B (Neither) to A (Both), rather than the gold D (Visual-only). Red marks incorrect predictions.

and test data are separated by source media. Appendix A details the construction, verification, partitioning, and semantic coverage.

Evaluation metrics. We report final-answer accuracy (AA), modality-judgment accuracy (JA), joint accuracy of answers and modality judgments (JAA), cross-modal hallucination rate (Hall.), and harmful flip rate (Flip). All five are percentages. AA scores the final option; sample-level JA requires all four candidate judgments in a detection question, or both modality judgments in a quartet input, to be correct. JAA requires both the final answer and the required judgments to be correct; the combined-state field is not scored separately. For these three correctness metrics, we use

$$m_{\text{Macro}} = \frac{m_A + m_V + m_Q}{3}, \tag{1}$$

where A, V, and Q denote audio detection, visual detection, and quartet evaluation. Judgments are scored as reported, not inferred from the final option. The AA-JAA gap measures correct answers accompanied by incorrect required judgments under this joint-output protocol.

Cross-modal hallucination rate. Hall. counts false-presence judgments for predefined absent targets with potentially misleading cues from the other modality. It averages the audio detection, visual detection, and quartet false-presence rates equally.

Harmful flip rate. Flip measures how often an initially correct absent judgment changes to present when only the other modality changes. Neither-to-Visual-only holds audio fixed and changes video; Neither-to-Audio-only holds video fixed and changes audio. Each directional rate conditions on initially correct absence judgments; Flip averages the two directions. Appendix C gives the full definitions, denominator conventions, and event counts.

3.2 Factorial Quartet Fine-Tuning

FQFT couples correctness of individual responses with factorial regularization of modality judgments across related examples (Figure 6). A training quartet groups diverse clips of the same queried sound and visual content into *Both*, *Audio-only*, *Visual-only*, and *Neither*. Training groups are defined by their audio and visual labels, allowing different clips for the same modality label within a quartet. Their contrasts regularize label-associated variation, including score differences between same-label clips; evaluation quartets reuse clips to measure changes when only the other modality varies.

Modality-judgment margins. Let x_{av} denote an audio-visual input, where $a = +1$ if the audio contains the queried sound and $a = -1$ otherwise, and $v = +1$ if the video shows the queried visual content and $v = -1$ otherwise. For queried modality $q \in \{A, V\}$, define

$$m^q_{av} = z_\theta(\text{present} \mid x_{av}, p_q) - z_\theta(\text{absent} \mid x_{av}, p_q), \tag{2}$$

where z_θ is the prediction logit and p_q queries the corresponding target without preceding generated modality judgments. Both audio and video remain available. A positive margin favors presence; a negative margin favors absence. These independent queries provide auxiliary regularization; rewards and evaluation use complete autoregressive responses.

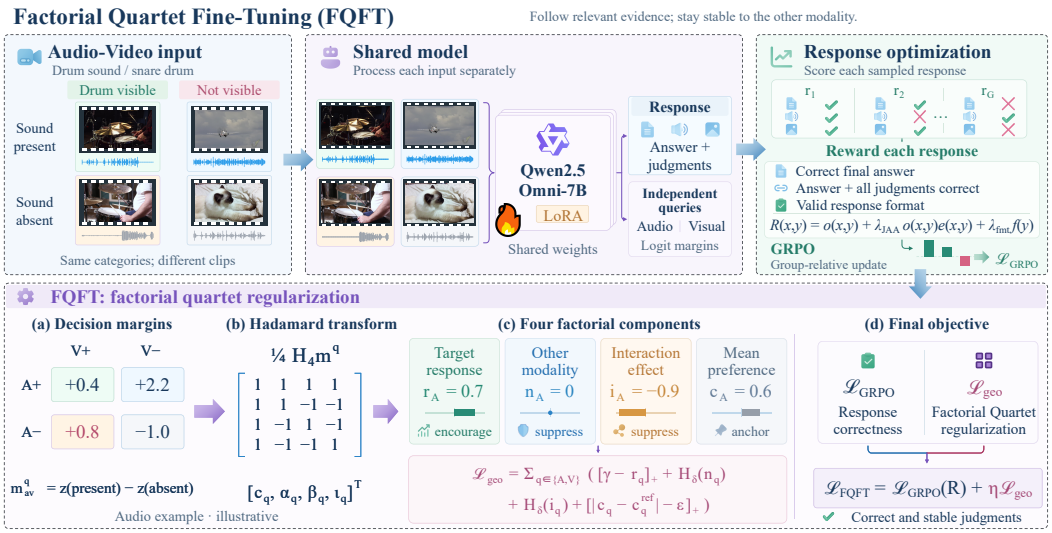


Figure 6: Factorial Quartet Fine-Tuning (FQFT). Training quartets group diverse clips of the same queried sound and visual content according to their audio and visual labels. The shared model generates complete responses for GRPO and independent modality-judgment margins for quartet regularization. (a-c) The Hadamard transform separates queried-modality, other-modality, interaction, and mean components. The illustrative audio margins have zero average other-modality contrast but nonzero interaction, exposing changes hidden by the average. (d) FQFT combines response-level correctness with queried-modality sensitivity, suppression of other-modality and interaction contrasts, and mean anchoring to the frozen SFT reference.

Factorial decomposition of judgment margins. Within each quartet, the four margins admit the exact factorial expansion

$$m^q_{av} = c_q + a\alpha_q + v\beta_q + av\iota_q, \tag{3}$$

with coefficients obtained by an orthogonal Hadamard transform:

$$c_q = \tfrac{1}{4}\sum_{a,v} m^q_{av}, \quad \alpha_q = \tfrac{1}{4}\sum_{a,v} a\, m^q_{av}, \quad \beta_q = \tfrac{1}{4}\sum_{a,v} v\, m^q_{av}, \quad \iota_q = \tfrac{1}{4}\sum_{a,v} av\, m^q_{av}. \tag{4}$$

Here c_q is the mean margin, α_q and β_q are the audio- and visual-label contrasts, and ι_q is their interaction coefficient. Reordering the coefficients by the queried modality gives

$$(r_A, n_A, i_A) = (\alpha_A, \beta_A, \iota_A), \qquad (r_V, n_V, i_V) = (\beta_V, \alpha_V, \iota_V). \tag{5}$$

The *queried-modality contrast* r_q measures whether average margins separate clips with and without the queried sound or visual content in the correct direction. The *other-modality contrast* n_q measures average variation associated with the accompanying modality's labels. The *interaction* i_q captures how this variation differs between the two groups defined by the queried modality.

For matched inputs, the largest margin change when the accompanying modality changes is $2(|n_q| + |i_q|)$; both matched changes vanish exactly when $n_q = i_q = 0$. This matched-input identity motivates a label-level consistency penalty for training quartets. Appendix D derives these identities and their relation to parallel matched comparisons.

Factorial quartet regularization. FQFT encourages the queried-modality contrast r_q to be at least γ while penalizing n_q and i_q. These contrasts are unchanged by a common shift of all four margins, so we additionally anchor the mean c_q to its frozen SFT reference value. With Huber penalty H_δ and $[u]_+ = \max(u, 0)$, the regularizer is

$$\mathcal{L}_{\text{geo}} = \sum_{q\in\{A,V\}} \left([\gamma - r_q]_+ + H_\delta(n_q) + H_\delta(i_q) + [|c_q - c_q^{\text{ref}}| - \epsilon]_+\right). \tag{6}$$

Here c_q^{ref} is the SFT model's mean margin on the same four training inputs. The terms strengthen the queried-modality contrast, penalize average and interaction contrasts, and limit changes in the overall preference for the present label. When $n_q = i_q = 0$, the margins for the two queried-modality labels are $c_q + r_q$ and $c_q - r_q$; they have the correct strict signs exactly when $r_q > |c_q|$. Response-level supervision promotes correct answers and judgments, while quartet regularization controls discrimination and variation of the directly queried margins.

Response-level optimization. We optimize complete responses with GRPO (Shao et al., 2024). For response y to input x, let $o(x, y)$ and $e(x, y)$ indicate final-answer and modality-judgment correctness, and let $f(y)$ indicate a valid response format. The reward is

$$R(x, y) = o(x, y) + \lambda_{\text{JAA}} o(x, y)e(x, y) + \lambda_{\text{fmt}} f(y). \tag{7}$$

The joint correctness term $o(x, y)e(x, y)$ rewards a response only when its final answer and every scored modality judgment are correct; the combined-state field is not scored separately. All policy-training variants share the answer and format terms. The full objective is

$$\mathcal{L}_{\text{FQFT}} = \mathcal{L}_{\text{GRPO}}(R) + \eta \mathcal{L}_{\text{geo}}. \tag{8}$$

The quartet regularizer applies only to complete training quartets; audio and visual detection questions use only the response-level objective. Appendix E details margin queries and optimization.

4 Experiments

The experiments examine the relationship between final-answer correctness, modality-judgment correctness, and cross-modal stability on AV-TRACE. Controlled comparisons with supervised fine-tuning and vanilla GRPO assess the effectiveness of FQFT on Qwen2.5-Omni-7B, complemented by evaluation on an adapted external benchmark.

4.1 Experimental Setup

Models and comparisons. Table 2 reports five released models (Xu et al., 2025a;b; Cui et al., 2026; Gemma Team, 2026) and five mitigation baselines. Our SFT baseline uses Qwen2.5-Omni-7B, whereas ACPO uses VideoLLaMA2.1-7B-AV. All training-based methods in our comparison use the same task-training data; their backbones may differ. To isolate objective-level differences, our Qwen2.5-Omni policy-training variants start from the same SFT checkpoint. MoD-DPO (Chaubey et al., 2026) is not included in the quantitative comparison because its official implementation was not publicly available at the time of our experiments.

Training and evaluation. All models receive the same evaluation questions and response schema through their respective input processors. We report task-macro AA, JA, JAA, and Hall., together with the mean directional Flip. Appendix E gives the training, input-processing, and optimization settings.

4.2 Comparative Experiments

Table 2: Results on AV-TRACE (%). Bold/underline mark best/second-best scores for released models and mitigation baselines separately. For Qwen2.5-Omni-7B, bold marks the best trained variant.

Model / method	Correctness			Cross-modal errors	
	AA↑	JA↑	JAA↑	Hall.↓	Flip↓
Native Omni					
Qwen2.5-Omni-7B	<u>56.99</u>	44.96	44.29	58.90	74.82
Qwen3-Omni-30B-A3B	56.42	<u>53.72</u>	<u>53.72</u>	**20.44**	**14.48**
MiniCPM-o-4.5	**67.80**	**55.14**	**55.10**	<u>26.78</u>	27.48
Unified multimodal					
Gemma-4-E4B	48.43	47.79	47.79	30.11	27.13
Gemma-4-12B	49.82	49.65	49.62	34.16	<u>23.53</u>
AVCD (Jung et al., 2025)	50.63	41.60	41.11	**21.55**	<u>27.09</u>
MAD (Chung et al., 2026)	**62.32**	**55.94**	**54.11**	43.79	35.96
RL-CoMM (Ye et al., 2026)	<u>57.38</u>	45.35	<u>44.71</u>	58.40	33.35
ACPO (Baid et al., 2026)	56.80	<u>52.23</u>	34.77	<u>29.80</u>	27.51
ASD (Jung et al., 2026)	45.68	38.83	30.49	33.48	**22.38**
Qwen2.5-Omni-7B					
+ SFT	77.56	76.32	76.29	20.49	4.08
+ SFT + Vanilla GRPO	83.03	81.44	81.43	14.23	2.40
+ SFT + FQFT (ours)	**85.75**	**84.47**	**84.39**	**10.32**	**1.95**

Answer-judgment accuracy gap. Qwen2.5-Omni's AA-JAA gap is 12.70 percentage points (56.99% versus 44.29%), reflecting correct answers with incorrect required modality judgments (Figure 7). Gemma-4-E4B and Gemma-4-12B have smaller gaps (0.64 and 0.20 points), but lower

JAA and higher Hall. and Flip than Qwen3-Omni. A smaller gap thus implies neither higher joint correctness nor greater cross-modal stability.

Modality-level correctness and cross-modal stability. Among the released models, Hall. ranges from 20.44% to 58.90% and Flip from 14.48% to 74.82%. The matched comparisons reveal initially correct absence judgments becoming false presence despite an unchanged queried-modality clip. MiniCPM-o-4.5 has higher JAA than Qwen3-Omni-30B-A3B (55.10% versus 53.72%) but a higher Flip rate (27.48% versus 14.48%; Figure 8). Flip conditions on each model's initially correct absence judgments and complements correctness by measuring harmful transitions. These results distinguish accuracy on individual inputs from stability under changes in the other modality.

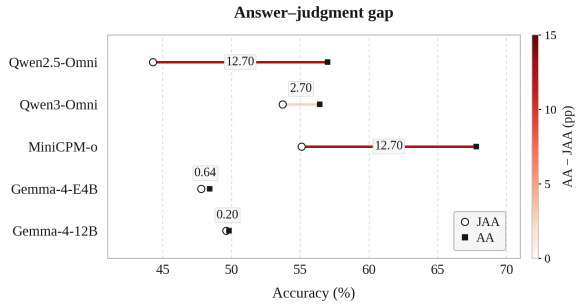


Figure 7: Answer-judgment gaps for five released models on AV-TRACE. Final-answer accuracy (AA, squares) and joint answer-judgment accuracy (JAA, circles); labels give AA-JAA gaps (pp): correct answers with incorrect required judgments. AA/JAA↑.

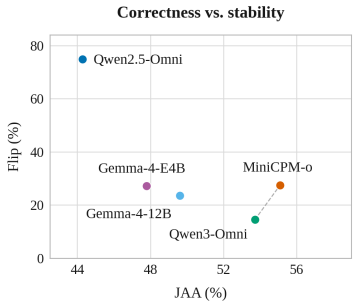


Figure 8: JAA and mean conditional Flip for five released models on AV-TRACE. MiniCPM-o has higher JAA and Flip than Qwen3-Omni. JAA↑, Flip↓.

Effectiveness of FQFT. On Qwen2.5-Omni-7B, FQFT improves all five metrics over SFT and vanilla GRPO (Table 2). Relative to vanilla GRPO from the same SFT initialization, JAA increases by 2.96 percentage points, while Hall. and Flip decrease by 3.91 and 0.45 points. The largest JAA gains over SFT occur in audio and visual detection (+11.70 and +11.40 points), compared with +1.19 points on quartet inputs (Appendix F, Table 5). The task-macro gain is thus primarily detection-driven, whereas lower Flip indicates fewer harmful transitions on matched inputs. FQFT reduces Flip in both directions; the remaining observed flips are visual-to-audio (Appendix F). Quartet recall is 93.23% for present audio targets and 92.18% for present visual targets, ruling out uniformly absent predictions.

4.3 Ablation Experiments

Table 3 crosses the joint answer-judgment reward and factorial quartet regularizer to assess separate and joint effects. All policy variants share the 2,500-update SFT checkpoint and use 3,500 updates with shared answer and format rewards. *Vanilla GRPO* uses only these terms; *GRPO + Joint reward* adds the joint-correctness bonus; *GRPO + Quartet* adds the quartet regularizer; FQFT uses both.

Table 3: Ablation on Qwen2.5-Omni-7B (%). All policy variants share the SFT checkpoint. Joint reward adds the joint-correctness bonus; Quartet adds the factorial regularizer; FQFT combines both. Bold/underline mark best/second-best; superscripts give changes from SFT (percentage points).

Training	Correctness			Cross-modal errors	
	AA↑	JA↑	JAA↑	Hall.↓	Flip↓
Qwen2.5-Omni-7B					
SFT	77.56	76.32	76.29	20.49	4.08
+ Vanilla GRPO	83.03$^{+5.47}$	81.44$^{+5.12}$	81.43$^{+5.14}$	14.23$^{-6.26}$	2.40$^{-1.68}$
+ GRPO + Joint reward	82.64$^{+5.08}$	81.39$^{+5.07}$	81.37$^{+5.08}$	14.58$^{-5.91}$	2.28$^{-1.80}$
+ GRPO + Quartet	<u>85.19</u>$^{+7.63}$	<u>83.30</u>$^{+6.98}$	<u>83.29</u>$^{+7.00}$	<u>10.88</u>$^{-9.61}$	**1.83**$^{-2.25}$
+ FQFT (ours)	**85.75**$^{+8.19}$	**84.47**$^{+8.15}$	**84.39**$^{+8.10}$	**10.32**$^{-10.17}$	<u>1.95</u>$^{-2.13}$

The joint reward alone does not improve joint accuracy over vanilla GRPO: JAA is 81.37% versus 81.43%, with slightly higher Hall. and lower Flip. Adding quartet regularization improves all five metrics both without and with the joint reward, increasing JAA by 1.86 and 3.02 points, respectively. Hall. falls by 3.35 and 4.26 points, and Flip by 0.57 and 0.33 points, in these two comparisons. The response schema, training data, answer/format rewards, and SFT initialization are shared, so these gains establish the quartet module's incremental contribution beyond the common structured-

output training setup. Target discrimination, within-label score consistency, and mean anchoring are constituent mechanisms of this module, evaluated jointly by the ablation. The complete-response results assess its downstream effects on correctness and harmful transitions (Appendix F).

Adding the joint reward to the quartet regularizer yields the highest AA, JA, and JAA and the lowest Hall. among the ablated variants. JAA increases from 83.29% to 84.39%, and Hall. decreases from 10.88% to 10.32%. The quartet-only variant retains the lower Flip rate (1.83% versus 1.95%). Thus, the joint reward strengthens response-level correctness when combined with quartet regularization, but does not further reduce harmful flips. Appendix F gives directional errors and positive-target recall.

4.4 Cross-Benchmark Evaluation

We evaluate the trained Qwen2.5-Omni variants on 6,408 AVHBench examples (Kim et al., 2025), adapting the prompts to request task-specific judgments and final answers. We denote this protocol by AVHBench* (Table 4); scores use our response format rather than the original benchmark protocol. Caption uses four-choice description selection and deterministic scoring. Correctness is averaged equally over Joint, AV Match, and Caption, whereas Hall. averages the two directional false-presence rates within Joint. AVHBench* has no shared-clip quartets and therefore does not support Flip. Appendix G details adaptation and scoring.

Table 4: Results on AVHBench* (%). Bold/underline rank released models and mitigation baselines separately; bold marks the best Qwen2.5-Omni trained variant.

Model / method	AA↑	JA↑	JAA↑	Hall.↓
Native Omni				
Qwen2.5-Omni-7B	**76.78**	44.77	44.61	84.86
Qwen3-Omni-30B-A3B	70.58	40.56	40.56	<u>26.80</u>
MiniCPM-o-4.5	<u>72.37</u>	39.25	39.24	31.81
Unified multimodal				
Gemma-4-E4B	68.23	<u>61.09</u>	<u>61.09</u>	35.62
Gemma-4-12B	68.23	**64.66**	**64.66**	**20.77**
AVCD (Jung et al., 2025)	52.29	61.59	49.56	87.82
MAD (Chung et al., 2026)	<u>73.20</u>	**71.61**	**71.22**	84.97
RL-CoMM (Ye et al., 2026)	**77.15**	45.08	44.92	85.04
ACPO (Baid et al., 2026)	53.59	57.65	39.69	**37.21**
ASD (Jung et al., 2026)	70.51	<u>64.73</u>	<u>62.14</u>	<u>64.02</u>
Qwen2.5-Omni-7B				
+ SFT	**76.85**	49.30	49.04	23.18
+ SFT + Vanilla GRPO	69.87	54.62	54.62	12.55
+ SFT + FQFT (ours)	75.51	**75.52**	**75.50**	**9.38**

FQFT achieves 75.50% JAA and 9.38% Hall., the best JAA and Hall. in Table 4. Its AA is 75.51%, compared with 76.85% for SFT, so the advantage concerns joint answer-judgment correctness rather than final-answer accuracy alone. Both policy variants start from the same Qwen2.5-Omni-7B checkpoint after SFT on 20K examples. Relative to SFT, FQFT increases JAA by 26.46 points and reduces Hall. by 13.80 points. The AA–JAA gap narrows from 27.81 points for SFT to 0.01 for FQFT, showing that correct answers more often coincide with correct task-specific judgments. FQFT improves JA and JAA by 20.90 and 20.88 percentage points and reduces Hall. by 3.17 points relative to vanilla GRPO. The paired gains in JA and JAA show that the transfer benefit extends to task-specific decisions, not only final-option selection, under the adapted protocol.

5 Conclusion

AV-TRACE combines modality-specific hallucination detection and cross-modal stability testing to distinguish correct final answers, correct modality judgments, and stable judgments across matched inputs. FQFT addresses this distinction by jointly optimizing answer-judgment correctness and the variation of modality judgments across four target-presence states. On Qwen2.5-Omni-7B, it improves joint answer-judgment accuracy and reduces cross-modal hallucination and harmful flip rates relative to SFT on AV-TRACE. The trained Qwen2.5-Omni model also achieves strong joint correctness and a low hallucination rate on AVHBench*. Together, these results support training and evaluating modality judgments for both correctness and selective sensitivity to the queried modality.

AI USE STATEMENT

We used GPT-5.6 for language editing, consistency checks, literature discovery, and feedback on methodological presentation and interpretation of results; all cited sources were verified by the authors. Qwen3.6-27B assisted preliminary screening of candidate data. The curated data and adapted task materials underwent human verification against the source media, with final inclusion decisions made by the authors. Generative AI was not used to generate experimental results. We reviewed all AI-assisted work and take responsibility for the paper's text, claims, results, and associated materials.

REFERENCES

A. Baid, Z. Xue, and K. Grauman. Don't let the video speak: Audio-contrastive preference optimization for audio-visual language models. In *Computer Vision – ECCV 2026*, volume 17076 of *Lecture Notes in Computer Science*, pp. 54–72, 2026.

A. Chaubey, J. Pang, and M. Soleymani. MoD-DPO: Towards mitigating cross-modal hallucinations in omni LLMs using modality decoupled preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18284–18294, 2026.

H. Chen, W. Xie, A. Vedaldi, and A. Zisserman. VGGSound: A large-scale audio-visual dataset. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725, 2020.

J. Chen, T. Zhang, S. Huang, Y. Niu, C. Sun, R. Zhang, G. Zhou, and L. Wen. OmniDPO: A preference optimization framework to address omni-modal hallucination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20172–20180, 2026.

Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, C. Zhou, and J. Zhou. Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

S. Chung, S. Y. Kim, Y. Chee, and Y. M. Ro. MAD: Modality-adaptive decoding for mitigating cross-modal hallucinations in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40175–40185, 2026.

J. Cui, B. Xu, C. Wang, T. Yu, W. Sun, Y. Xu, T. Wang, Z. He, W. Ma, T. Cai, J. Gui, L. Zhang, X. Sun, F. Huang, M. Chen, Z. Lin, H. Liu, Q. Gui, Q. Han, Y. Wen, H. Liu, R. Wang, Y. Zhang, H. Wei, C. Chen, Y. Li, K. Fang, J. Zhou, Y. Li, G. Zeng, C. Xiao, Y. Lin, X. Han, M. Sun, Z. Liu, and Y. Yao. MiniCPM-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Z. Dong, J. Tang, Z. Lei, Z. Cao, Z. Zhang, Y. Wang, S. Li, X. Wang, B. Peng, and J. Liu. OmniHalluc-L: Counterfactual benchmarking and modality-perturbation reliability calibration for long-form omni hallucination. *arXiv preprint arXiv:2606.03614*, 2026.

Gemma Team. Gemma 4 Technical Report. *arXiv preprint arXiv:2607.02770*, 2026.

Y. Gong, Y.-A. Chung, and J. Glass. AST: Audio spectrogram transformer. In *Proceedings of Interspeech 2021*, pp. 571–575, 2021.

T. Guan, F. Liu, X. Wu, R. Xian, Z. Li, X. Liu, X. Wang, L. Chen, F. Huang, Y. Yacoob, D. Manocha, and T. Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, 2024.

C. Jung, Y. Jang, and J. S. Chung. AVCD: Mitigating hallucinations in audio-visual large language models through contrastive decoding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 63143–63174, 2025.

J. Jung, C. Jung, J.-H. Kim, and J. S. Chung. Probing cross-modal information hubs in audio-visual LLMs. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

S.-B. Kim, H.-B. Oh, J.-M. Lee, A. Senocak, J. S. Chung, and T.-H. Oh. AVHBench: A cross-modal hallucination benchmark for audio-visual large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

S. Leng, Y. Xing, Z. Cheng, Y. Zhou, H. Zhang, X. Li, D. Zhao, S. Lu, C. Miao, and L. Bing. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

J. Li, X. Cheng, X. Zhao, J.-Y. Nie, and J.-R. Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6449–6464, 2023a.

X. Li, Y. Lu, C. Cui, T. Wang, Z. Huang, Y. Jiang, L. Zhou, L. Wang, and J. Dang. Separate first, fuse later: Mitigating cross-modal interference in audio-visual LLMs reasoning with modality-specific chain-of-thought. *arXiv preprint arXiv:2605.09906*, 2026.

Y. Li, Y. Du, K. Zhou, J. Wang, X. Zhao, and J.-R. Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 292–305, 2023b.

J. Liu, S. Chen, X. He, L. Guo, X. Zhu, W. Wang, and J. Tang. VALOR: Vision-audio-language omni-perception pretraining model and dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 47(2):708–724, 2025.

P. Manakul, A. Liusie, and M. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9004–9017, 2023.

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pp. 28492–28518, 2023.

Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, and D. Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

A. Sudheer, D. Li, O. Elbanna, I. Kodarapu, A. Bahuguna, and V. Sharma. Compositional failure in audio-visual LLMs: Late-layer prior dominance under cross-modal conflict. In *Second Workshop on Compositional Learning at ICML 2026*, 2026.

G. Sun, W. Yu, C. Tang, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, Y. Wang, and C. Zhang. video-SALMONN: Speech-enhanced audio-visual large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235, pp. 47198–47217, 2024.

X. Wen, W. J. Mo, X. Fu, R. Cai, T. Zhu, W. Li, Y. Xie, M. Chen, and P. Qi. When vision speaks for sound. *arXiv preprint arXiv:2605.16403*, 2026.

Y. Wu, K. Chen, T. Zhang, Y. Hui, M. Nezhurina, T. Berg-Kirkpatrick, and S. Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang, B. Zhang, X. Wang, Y. Chu, and J. Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

J. Xu, Z. Guo, H. Hu, Y. Chu, X. Wang, J. He, Y. Wang, X. Shi, T. He, X. Zhu, Y. Lv, Y. Wang, D. Guo, H. Wang, L. Ma, P. Zhang, X. Zhang, H. Hao, Z. Guo, B. Yang, B. Zhang, Z. Ma, X. Wei, S. Bai, K. Chen, X. Liu, P. Wang, M. Yang, D. Liu, X. Ren, B. Zheng, R. Men, F. Zhou, B. Yu, J. Yang, L. Yu, J. Zhou, and J. Lin. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Q. Ye, W. Zeng, M. Liu, J. Zhang, Y. Hu, Z. Yu, and Y. Zhou. When eyes and ears disagree: Can MLLMs discern audio-visual confusion? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11955–11963, 2026.

Z. Zhang, W. Zhou, J. Zhao, and H. Li. Robust multimodal large language models against modality conflict. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267, pp. 77233–77253, 2025.

A Dataset Construction Details

A.1 Construction and coverage

Construction and verification. AV-TRACE draws on VGGSound (Chen et al., 2020) and VALOR (Liu et al., 2025). Candidate sounds and visual content are drawn from source annotations and verified separately in audio and video. Audio verification combines speech recognition, sound-event classification, audio-text similarity, and captioning. Visual verification uses sampled frames; negative labels are additionally checked against the original videos, and content labeled as visible must remain visible in the frames supplied to the model. Labels require support from the corresponding modality, and unresolved cases are excluded. Clips without the queried sound or visual content retain other meaningful audio or video content.

Verified clips are assembled into detection questions and matched quartets. Two annotators independently check the source media, target labels, answer options, and shared-clip consistency, with a third reviewer adjudicating disagreements. Only unambiguous examples that pass quality review are retained, following the procedure below. After screening and human verification, the curated data are partitioned into training, validation, and test sets; source clips remain within a single partition, as do all four members of each evaluation quartet. This controls overlap within our constructed data, but not possible exposure to public source clips during model pretraining. The following paragraphs give annotation and partitioning details. The benchmark covers 399 audio target labels and 256 visual target labels spanning people, animals, musical instruments, machines, and everyday events or objects (Figure 2).

A.2 Annotation and partitioning details

Initial screening. We first construct a pool of approximately 180K candidate examples from the collected multimodal data. Initial screening removes corrupted or incomplete media, insufficient task-relevant evidence, duplicate content, invalid modality combinations, and examples that do not support reliable annotation. Approximately 130K candidates pass screening and enter human verification. This reduction reflects data curation, whereas the disagreement and audit rates below measure annotation quality.

Annotation models. Audio-MoE combines four pretrained components: Whisper large-v3-turbo provides a transcript (Radford et al., 2023); AST AudioSet provides sound-event scores (Gong et al., 2021); CLAP HTSAT-fused compares target aliases against control descriptions (Wu et al., 2023); and Qwen2-Audio-7B-Instruct produces an audio-only caption (Chu et al., 2024). Evidence fusion considers event agreement, target-control similarity, caption support, and media quality.

Qwen3.6-27B reviews the resulting audio evidence with source labels and fusion verdicts hidden, and independently assesses visual targets from sampled frames. This model pre-review covers approximately 80% of candidate samples and helps identify low-quality or potentially ambiguous cases. Its outputs serve as auxiliary screening signals rather than final annotation labels. Clips with and without the target undergo the same verification criteria. Model pre-review assists the subsequent human review and does not replace it.

Human verification. Data annotation follows a *model pre-review + 2+1 human review* procedure.

1. **Independent annotation.** Two annotators independently apply a unified annotation guideline, without access to each other's first-pass judgments. They check sound presence against the audio and visual presence against the video, without inferring one modality from the other. They verify the target labels, positive and negative candidates, cross-modal distractor, and final option mapping. Visual absence is checked against the original video rather than sampled frames alone; positive visual content must remain visible in the frames supplied to the model. For quartets, the review additionally checks label consistency, answer mappings, and the shared-clip structure across all four conditions. Examples with agreed labels are provisionally accepted.
2. **Third-reviewer adjudication.** When the two annotators disagree, a third annotator independently re-examines the disputed example using the same guidelines and source media. Examples that do not satisfy the quality requirements are excluded.

3. **Secondary review and retention.** Ambiguous examples that may admit a definite judgment after further inspection enter a secondary review using the complete available context. They are retained only if this review reaches a clear conclusion and the example passes the quality checks; otherwise they are excluded.

The final inclusion criterion is an unambiguous label supported by the corresponding modality and a completed quality review, rather than automatic acceptance by majority vote.

Agreement and adjudication statistics. The two annotators initially disagree on 8.7% of all doubly annotated examples. Every disputed example is submitted to the third annotator. The adjudication rate, measured over the same doubly annotated set, is therefore also 8.7%, while adjudication coverage among disputed examples is 100%. Each disagreement is reviewed before the example is retained, relabeled, or excluded.

Independent quality audit. After annotation and adjudication, we independently audit 1,000 examples selected by category-stratified random sampling from the retained data. Reviewers did not participate in the original annotation of the examples they audit. The audit identifies 14 annotation errors, giving an error rate of 1.4% among the audited examples. This post-adjudication audit examines residual errors beyond the initial disagreements, including errors in examples initially accepted by both annotators. We also examine audit outcomes by annotation category to identify category-specific error patterns.

Data partitioning. Partitioning follows completion of annotation and quality control. Sampling is randomized within annotation categories, with category-level quotas limiting the dominance of frequent categories and maintaining representation across training and benchmark subsets. Source-media grouping constrains this sampling: examples sharing a source identifier remain in the same training, validation, or test partition, and all four members of an evaluation quartet are assigned jointly. Within the training partition, the SFT and reinforcement-learning subsets contain 20K and 28K task inputs, respectively, with category composition controlled separately for each stage. Together, these constitute 48K training task inputs across the two stages; the stages may share media. Source-grouped partitioning prevents overlap among our constructed training, validation, and test media, although it cannot establish whether publicly available source clips were encountered during model pretraining.

Benchmark counting units. The test set contains 7,344 task inputs: 1,000 audio hallucination detection questions, 1,000 visual hallucination detection questions, and 5,344 inputs organized into 1,336 matched quartets. Each quartet combines two audio clips and two video clips into four audio-video pairings. At the group level, the 2,000 standalone detection questions and 1,336 quartets constitute 3,336 evaluation units. These units describe the evaluation structure rather than establish statistical independence. Source-media identity is tracked separately: audio and video derived from the same underlying source share one source identity, and shared source identifiers define dependencies across evaluation units.

B Examples of the Two Evaluations

Figures 4 and 5 show test-set media, verified target labels, and illustrative model predictions. The displayed questions are abridged; candidate names, option mappings, and modality judgments are retained.

B.1 Modality-Specific Hallucination Detection

Figure 4 illustrates visual-to-audio and audio-to-visual distractors. In the first example, the visible tabla suggests a sound that is absent from the queried audio. In the second, the audible printer analogously suggests an entity that is absent from the queried video.

B.2 Cross-Modal Stability Testing

Figure 5 queries *dog barking* in audio and *dog* in video across a complete quartet, with other audible and visible content retained in the clips without the respective targets. With the same non-barking audio, the illustrated prediction reports the audio target as absent in Neither but present in Visual-only

after only the video changes. Its final answer correspondingly changes from Neither to Both, although the gold audio judgment remains absent in both conditions. The harmful flip is therefore visible in both the explicit modality judgment and the predicted final state.

B.3 Complete input prompts and reference outputs

The following prompts and reference outputs are reproduced from the evaluation manifest, without abridging the instructions or output schema. The examples match the targets and option mappings in Figures 4-5. The outputs below are *gold reference responses*, not recorded model predictions; the erroneous predictions in the figures remain illustrative. The markers <video><audio> indicate the media attachments supplied alongside the text through the model-specific processor. Candidate names, target names, and option mappings are instantiated per question. Within a matched quartet, the same prompt and option mapping are used with all four audio-visual combinations.

```
<video><audio>
Which one of these fine-grained sound events is actually audible?

A. "cat purring"
B. "playing tabla"
C. "dog barking"
D. "bird squawking"

For each named candidate, decide independently whether it is audible; then
  ↪ choose the one present candidate. Return exactly the structure below and do
  ↪ not add reasoning.

[Audio candidate decisions]
"cat purring": <present or absent>
"playing tabla": <present or absent>
"dog barking": <present or absent>
"bird squawking": <present or absent>

[Final answer]
Answer: <A, B, C, or D>
```

Listing 1: AV-TRACE audio detection: complete input prompt.

```
<video><audio>
Which one of these fine-grained visual candidates is actually visible?

A. "dog"
B. "pig"
C. "printer"
D. "man working on car"

For each named candidate, decide independently whether it is visible; then
  ↪ choose the one present candidate. Return exactly the structure below and do
  ↪ not add reasoning.

[Visual candidate decisions]
"dog": <present or absent>
"pig": <present or absent>
"printer": <present or absent>
"man working on car": <present or absent>

[Final answer]
Answer: <A, B, C, or D>
```

Listing 2: AV-TRACE visual detection: complete input prompt.

Audio detection: reference output.

```
[Audio candidate decisions]
"cat purring": absent
"playing tabla": absent
"dog barking": present
"bird squawking": absent

[Final answer]
Answer: C
```

Visual detection: reference output.

```
[Visual candidate decisions]
"dog": present
"pig": absent
"printer": absent
"man working on car": absent

[Final answer]
Answer: A
```

Matched quartet: shared input text.

Both modalities are supplied in every condition; only the attached audio and video differ.

```
<video><audio>
Evaluate the sound event and visual entity named in the options. Use only the
  ↪ audio for the sound event and only the video for the visual entity; do not
  ↪ infer one modality from the other.

Which option is supported by the audio and video?

A. The target sound "dog barking" is audible and the target visual "dog" is
  ↪ visible.
B. Neither target is supported.
C. Only the target sound "dog barking" is supported.
D. Only the target visual "dog" is supported.

Return exactly the following structure. Do not add reasoning or extra sections.

[Decisions]
Audio target: <present or absent>
Visual target: <present or absent>
Joint state: <both_supported, audio_only, visual_only, or neither_supported>

[Final answer]
Answer: <A, B, C, or D>
```

Listing 3: AV-TRACE matched quartet: shared input prompt.

Matched quartet: four reference outputs.

Both

```
[Decisions]
Audio target: present
Visual target: present
Joint state: both_supported

[Final answer]
Answer: A
```

Audio-only

```
[Decisions]
Audio target: present
Visual target: absent
Joint state: audio_only

[Final answer]
Answer: C
```

Visual-only

```
[Decisions]
Audio target: absent
Visual target: present
Joint state: visual_only

[Final answer]
Answer: D
```

Neither

```
[Decisions]
Audio target: absent
Visual target: absent
Joint state: neither_supported

[Final answer]
Answer: B
```

C SUPPLEMENTARY SCORING CONVENTIONS

C.1 FULL DEFINITIONS OF THE CROSS-MODAL ERROR METRICS

Cross-modal hallucination rate. Hall. counts false-presence judgments for predefined absent targets with potentially misleading cues from the other modality. For detection question i, let $\hat{s}_{qi} \in \{-1, +1\}$ be the judgment of the designated cross-modal distractor: absent or present. With $q \in \{A, V\}$ denoting the queried modality and $\mathcal{C}_q$ its 1,000 detection questions,

$$H_q = \frac{100}{|\mathcal{C}_q|} \sum_{i \in \mathcal{C}_q} \mathbf{1}[\hat{s}_{qi} = +1]. \tag{9}$$

Only the designated distractor contributes to this rate; a false-presence judgment for another absent candidate still makes JA and JAA incorrect. For quartet evaluation, H_Q pools audio false presence in Visual-only and visual false presence in Audio-only, giving 2,672 eligible judgments. Both and Neither remain part of the correctness metrics but do not meet this absent-target/present-cue condition. The aggregate rate is

$$\text{Hall.} = \frac{H_A + H_V + H_Q}{3}. \tag{10}$$

Harmful flip rate. Flip measures how often an initially correct absent judgment changes to present when only the other modality changes. Neither-to-Visual-only holds audio fixed and changes video; Neither-to-Audio-only holds video fixed and changes audio. Both inputs in each pair have the same gold judgment, absent, for the queried modality. Let $\hat{s}^0_{qj}$ and $\hat{s}^1_{qj}$ denote judgments before and after the change, and let $\mathcal{P}_q$ be the fixed set of matched pairs for queried modality q. Then

$$F_q = 100 \, \frac{\sum_{j \in \mathcal{P}_q} \mathbf{1}[\hat{s}^0_{qj} = -1, \, \hat{s}^1_{qj} = +1]}{\sum_{j \in \mathcal{P}_q} \mathbf{1}[\hat{s}^0_{qj} = -1]}, \qquad \text{Flip} = \frac{F_A + F_V}{2}. \tag{11}$$

Each direction contains one matched pair per test quartet, giving 1,336 evaluated pairs. Flip measures the risk that an initially correct judgment becomes incorrect after the other modality changes. Pairs whose initial judgment is already incorrect are therefore excluded from this conditional denominator, although they remain part of the full evaluation set. The denominator is thus the model- and direction-specific number of initially correct absence judgments, rather than the total number of evaluated pairs. A zero denominator yields "–", and the mean is reported only when both directions are defined. All four states contribute to AA, JA, JAA, and the factorial margin analysis; Flip specifically measures newly introduced false presence along the two absent-target transitions. Cross-model Flip comparisons describe conditional error risks on each model's initially correct subset, rather than on a common subset of pairs. They are interpreted alongside correctness scores; positive-target recall separately measures single-input omissions, not paired transitions. The scoring conventions and event counts are given below.

C.2 SCORING CONVENTIONS AND EVENT COUNTS

AV-TRACE correctness. Audio detection and visual detection implement modality-specific hallucination detection; quartet evaluation implements cross-modal stability testing. JA assigns one correctness indicator to each input, requiring all task-specific judgments to be correct, rather than averaging individual judgment indicators. For every task,

$$\text{JAA} \leq \min(\text{AA}, \text{JA}),$$

and the same relation holds for the equal-weight task-macro averages. Full-test scoring retains every example and evaluates the final answer and modality-judgment fields separately. The AA-JAA gap is a response-level diagnostic of the jointly elicited fields, rather than a measure of unobserved internal judgments. Equal task weighting describes task-balanced performance rather than a sample-weighted average over all inputs.

Hallucination rate. Hall. uses fixed eligible target sets: 1,000 designated cross-modal distractors in each detection task and 2,672 absent-target judgments in quartet evaluation. Only a predicted presence contributes a false positive. Hall. measures false presence on the designated cross-modal distractors, whereas JA and JAA also penalize other incorrect required judgments.

Let $\bar{q}$ denote the other modality and define

$$\mathcal{Q}^{\text{conf}} = \{(i, q) : i \text{ is a quartet input}, \ s_{qi} = -1, \ s_{\bar{q}i} = +1\}.$$

The quartet component of the headline hallucination rate is

$$H_Q = \frac{100}{|\mathcal{Q}^{\text{conf}}|} \sum_{(i,q)\in\mathcal{Q}^{\text{conf}}} \mathbf{1}[\hat{s}_{qi} = +1]. \tag{12}$$

This includes one eligible absent-target judgment from every Visual-only example and one from every Audio-only example, with none from Both or Neither.

For SFT, the detection tasks contain 300 audio false positives and 223 visual false positives, each out of 1,000 questions, giving 30.00% and 22.30%. Its quartet false-positive counts are 208 in the visual-to-audio direction and 37 in the audio-to-visual direction, totaling 245 out of 2,672 eligible judgments and giving 9.17%. Averaging the three unrounded task rates yields 20.49% macro Hall.

For FQFT, the corresponding counts are 144 and 98 false positives, each out of 1,000 questions, giving 14.40% and 9.80%. In quartet evaluation, Visual-only audio judgments contribute 153 false positives and 1,183 true negatives, while Audio-only visual judgments contribute 28 false positives and 1,308 true negatives. Pooling these counts gives 181 false-presence judgments out of 2,672 eligible judgments, corresponding to 6.77% quartet Hall. Averaging the three unrounded task rates yields 10.32% macro Hall.

Conditional harmful flips. Each directional denominator counts initially correct absence judgments, as defined in Equation 11. For FQFT, 1,207 of the 1,336 visual-to-audio pairs have an initially correct audio-absence judgment, and 47 of these judgments become false presence after the video changes. In the audio-to-visual direction, 1,274 pairs have an initially correct visual-absence judgment, and none of these judgments becomes false presence after the audio changes. The resulting directional rates are 3.89% and 0.00%, respectively. Averaging the unrounded directional rates gives 1.95% Flip. Hall. counts false presence at the changed input, whereas Flip counts newly introduced errors; an incorrect judgment that persists across the change contributes to Hall. but not to Flip.

Positive-target judgments. For FQFT, recall on present targets in the quartet inputs is 93.23% for audio and 92.18% for vision, corresponding to omission rates of 6.77% and 7.82%. These recalls rule out a constant-absent prediction strategy.

D Factorial Decomposition and Matched Comparisons

Interpretation of training contrasts. Training quartets share target definitions and cover all four presence states, while allowing distinct media clips under the same label. The objective couples queried-label separation through r_q with group-level score consistency through n_q and i_q. In these groups, n_q and i_q capture other-label and interaction contrasts together with content- or difficulty-dependent variation between same-label queried-modality clips. Such clip variation can yield nonzero contrasts even for a judgment that uses only the queried modality. The training interaction thus characterizes non-additivity across labeled examples. Matched evaluation quartets provide the complementary controlled comparison: clip reuse holds queried-modality content fixed along each parallel edge, so the contrasts measure changes under replacement of the accompanying modality alone.

D.1 Matched margin changes

To relate this decomposition to cross-modal stability, consider two audio judgments that share the same audio clip but use different videos in a matched evaluation quartet:

$$m^A_{a,+} - m^A_{a,-} = 2(\beta_A + a\iota_A) = 2(n_A + ai_A). \tag{13}$$

This change combines the average other-modality contrast with an interaction component that depends on the audio label. When $n_A = 0$ and $i_A \neq 0$, the changes are equal and opposite for audio clips with and without the queried sound. More generally, let $\Delta_{q,+}$ and $\Delta_{q,-}$ denote the corresponding changes for queried-modality labels $+1$ and -1, respectively. Then

$$\Delta_{q,\pm} = 2(n_q \pm i_q), \qquad \max\{|\Delta_{q,+}|, |\Delta_{q,-}|\} = 2(|n_q| + |i_q|). \tag{14}$$

The largest matched margin change is therefore determined by both $|n_q|$ and $|i_q|$, and both changes vanish exactly when $n_q = i_q = 0$. FQFT uses these components to jointly regularize average and interaction contrasts within each training quartet. The derivation and the relation to parallel matched comparisons follow below.

D.2 Identification and correctness

Proposition 1 (Identification from a complete quartet)**.** *Without structural assumptions on the response surface, the complete* 2×2 *quartet is necessary and sufficient to identify* $(c_q, \alpha_q, \beta_q, \iota_q)$.

For a fixed queried modality, order the four margins as

$$\mathbf{m} = (m_{++}, m_{+-}, m_{-+}, m_{--})^\top$$

and the coefficients as

$$\mathbf{w} = (c, \alpha, \beta, \iota)^\top.$$

Equation 3 gives

$$\mathbf{m} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \mathbf{w} = H\mathbf{w}.$$

The Hadamard matrix satisfies

$$H^\top H = 4I,$$

so

$$\mathbf{w} = \frac{1}{4} H^\top \mathbf{m}.$$

The complete quartet is therefore sufficient to recover all four components.

To show necessity without structural assumptions, omit any one condition. The remaining observation matrix has rank at most three and therefore a nontrivial null space. At least two distinct coefficient vectors can consequently produce the same observed margins.

For example,

$$m_{++} - m_{-+} = 2(\alpha + \iota).$$

For any scalar t, replacing

$$(\alpha, \iota) \quad \text{with} \quad (\alpha + t, \iota - t)$$

leaves this matched contrast unchanged. Thus, target-associated variation and interaction cannot be separately identified from one edge alone. Parallel edges recover their sum and difference, while the remaining orthogonal contrasts recover the mean and other-modality component. The argument applies independently to both queried modalities.

Relation to matched comparisons. Let $\Delta_{q,+}$ and $\Delta_{q,-}$ denote other-modality contrasts when the queried target is present and absent, respectively. They satisfy

$$\Delta_{q,+} = 2(n_q + i_q), \qquad \Delta_{q,-} = 2(n_q - i_q).$$

Within the quadratic region of the Huber penalty,

$$H_\delta(n_q) + H_\delta(i_q) = \frac{\Delta_{q,+}^2 + \Delta_{q,-}^2}{16}.$$

The factorial components express the same four-condition variation as the two parallel contrasts, making their average and interaction contributions explicit. Within the quadratic regime, this yields a scaled sum of squared edge differences. Outside that regime, component-wise Huber penalties apply robustness to the average and interaction contrasts separately. Together with queried-label separation and mean anchoring, these terms form the auxiliary objective evaluated by the Quartet ablation in Table 3.

Stability and correctness. Write the queried-modality label as s and the other-modality label as t. Reordering Equation 3 gives

$$m^q(s, t) = c_q + s r_q + t n_q + s t i_q.$$

The two other-modality differences are $2(n_q+i_q)$ and $2(n_q-i_q)$. For real u, v, $\max(|u+v|, |u-v|) = |u| + |v|$, which proves Equation 14. Consequently, invariance of both matched margins is equivalent to $n_q = i_q = 0$. Correct sign decisions additionally require $s\, m^q(s,t) > 0$ in every condition. Since

$$\min_{s,t} s\, m^q(s,t) \geq r_q - |c_q| - |n_q| - |i_q|,$$

$r_q > |c_q| + |n_q| + |i_q|$ is a sufficient condition for all four decisions to be correct. Under exact invariance, this reduces to the necessary and sufficient condition $r_q > |c_q|$. The bound couples discrimination and invariance: with the other components fixed, increasing r_q strengthens correct-sign separation, while decreasing $|n_q|+|i_q|$ reduces the largest matched shift. For an initially negative margin, a harmful positive shift must cross zero; a larger initial distance below zero and a smaller shift provide complementary routes to avoiding this sign change. These statements concern the four directly queried margins; complete-response correctness is evaluated separately.

E Additional Experimental Details

Experimental environment and hyperparameters. Training experiments use eight NVIDIA GeForce RTX 4090 GPUs with 48 GB of memory per GPU. The software environment uses Python 3.14, CUDA 12.6, PyTorch 2.12.0, ms-swift 4.3.2, and FlashAttention 2.8.3. For Qwen2.5-Omni-7B, SFT uses a learning rate of 2×10^{-5}, a global batch size of eight examples, a maximum sequence length of 8,192 tokens, and seed 1024. Policy optimization uses a learning rate of 3×10^{-7}, a KL coefficient of 0.012, four completions per input, and a maximum completion length of 512 tokens. Both stages update language-side attention-projection LoRA modules with rank 16, scaling factor 32, and dropout 0.05, while keeping the perceptual encoders and multimodal aligner frozen. The reward coefficients are $\lambda_{\mathrm{JAA}} = 0.5$ and $\lambda_{\mathrm{fmt}} = 0.1$; quartet regularization uses the heuristically chosen weight $\eta = 0.005$, with $\gamma = 0.50$, $\epsilon = 0.10$, and $\delta = 1$. These values specify a fixed training protocol: the controlled ablations keep common hyperparameters unchanged and differ in the enabled objective terms.

Training and evaluation. Following the screening, human verification, and source-media partitioning described in Appendix A, the training partition supplies 20K examples for supervised fine-tuning and 28K examples for reinforcement learning. Validation contains 600 audio detection questions, 600 visual detection questions, and 400 quartets, totaling 2,800 inputs. Training stages may share media; validation and test media follow the held-out partitioning protocol in Section 3.1. All models receive the same evaluation questions and response schema through their respective input processors. We report task-macro AA, JA, JAA, and Hall., together with the mean directional Flip. Input processing and optimization settings are given below; Appendix B.3 provides complete example prompts and reference responses.

Baseline backbones. MAD, RL-CoMM, and ASD use Qwen2.5-Omni-7B-Instruct, whereas ACPO and AVCD use VideoLLaMA2.1-7B-AV. Our Qwen2.5-Omni SFT and policy-training comparisons also use the Instruct backbone.

Inputs and decision margins. Qwen2.5-Omni samples video at 2 fps, capped at eight frames per input, and receives separate audio resampled to 16 kHz through its released processor. Its SFT and FQFT evaluations use the same test inputs and greedy decoding (do_sample=False), with at most 512 generated tokens.

Decision margins are measured at Audio target: or Visual target:, without preceding generated judgments about the other modality. The factorial regularizer acts on these directly queried scores, while JA, JAA, Hall., and Flip are measured from judgments generated in complete responses. The selected tokenizer represents present and absent as distinct single tokens.

Independent margin-query format. Each query retains the original quartet question and both media inputs. After the processor renders the chat template with its assistant-generation marker, the audio query appends two newline characters and the following prefix:

```
[Decisions]
Audio target:
```

The visual query instead appends two newline characters and:

```
[Decisions]
Visual target:
```

The margin is read at the next-token position for the distinct single-token strings ” present” and ” absent”, including their leading spaces. No generated judgment about the other modality precedes either prefix. The difference between these two log-probabilities equals the logit difference in Equation 2. The independent visual query starts without an audio judgment, whereas the complete response generates its visual judgment after the audio judgment. The independent query conditions on the media, question, and a fixed prefix, whereas complete-response generation additionally conditions on preceding output tokens. FQFT uses the auxiliary margins to regularize target discrimination and score consistency, and the joint reward to supervise answer-judgment correctness in complete responses. These objectives update the shared model in their respective conditioning contexts. The margin identities in Appendix D apply to the auxiliary queries, and all reported performance scores are computed from complete responses.

Policy optimization. GRPO uses one sequence-level advantage per completion and KL regularization against the frozen initialization. The four completions sampled for one input form a GRPO reward group, distinct from the four inputs in a training quartet.

The Huber penalty in Equation 6 is

$$H_\delta(u) = \begin{cases} \frac{1}{2}u^2, & |u| \le \delta, \\ \delta(|u| - \frac{1}{2}\delta), & |u| > \delta. \end{cases}$$

Each run uses a frozen copy of its initialization for KL regularization; the Quartet variant and FQFT additionally use it for mean anchoring.

Quartet regularization is computed only for complete four-condition groups.

For each input, standard GRPO normalizes scalar rewards across its G sampled completions:

$$\hat{A}_k = \frac{R(x, y_k) - \bar{R}_x}{\mathrm{std}_x(R) + \varepsilon_{\mathrm{num}}}. \tag{15}$$

The same sequence-level advantage $\hat{A}_k$ is used at every completion token in the clipped policy objective. The four completions of one input define the GRPO reward group; four inputs covering the target-presence states form a separate training quartet.

Role of the joint reward. The joint bonus distinguishes answer-correct completions according to judgment correctness. If all answer-correct completions in a group also have correct judgments and the format reward is constant, adding the bonus is a positive affine transformation of the original rewards. Group normalization removes this transformation apart from the numerical stabilizer. The bonus therefore supplies a distinct relative reward signal when answer-correct completions differ in judgment correctness.

Overall ablation protocol. Vanilla GRPO, GRPO + Joint reward, GRPO + Quartet, and FQFT start from the same supervised checkpoint after 2,500 SFT updates and use 3,500 policy updates. The SFT baseline is the shared supervised checkpoint without policy optimization.

All policy-training variants share the answer and format reward

$$o(x, y) + \lambda_{\mathrm{fmt}} f(y).$$

The format-reward coefficient is 0.1, as specified above. The Joint reward and Quartet variants additionally use, respectively, the joint answer-judgment reward term in Equation 7 and the four-condition regularizer in Equation 6; FQFT uses both with the same coefficients. All variants use the sequence-level advantages in Equation 15 and are evaluated using the same AA, JA, JAA, Hall., and Flip definitions as the main comparison. The shared update count matches optimization length; the Quartet variant and FQFT additionally compute independent-query margins and reference margins for regularization.

F Detailed Diagnostic Results

The Qwen2.5-Omni FQFT result reported in the main paper uses the final checkpoint after 3,500 policy updates. The following breakdowns complement the headline metrics by separating task-level correctness from directional cross-modal errors. Figures 7 and 8 visualize the separation between answer correctness, joint answer-judgment correctness, and cross-modal stability across the five released models in Table 2.

Task-wise correctness. FQFT improves JAA over SFT by 11.70 points on audio detection and 11.40 points on visual detection, compared with 1.19 points on quartet inputs. Table 5 also reports FQFT AA and JA for each task. Under equal task weighting, the two detection tasks account for most of the task-macro JAA gain. Detection and quartet tasks share the trained parameters and response-level objective, while only complete quartets receive the additional margin regularizer. The quartet scores complement these task-balanced gains by assessing correctness and harmful transitions on matched inputs.

Table 5: Task-wise correctness (%). SFT reports JAA; FQFT reports AA, JA, and JAA.

Task	SFT	FQFT			ΔJAA (pp)
	JAA ↑	AA ↑	JA ↑	JAA ↑	
Audio detection	66.50	80.00	78.30	**78.20**	+11.70
Visual detection	77.00	90.70	88.50	**88.40**	+11.40
Quartet evaluation	85.37	86.56	86.60	**86.56**	+1.19
Task macro	76.29	**85.75**	**84.47**	**84.39**	+8.10

Directional cross-modal errors. Table 6 separates detection false presence, quartet false presence, and harmful flips by interference direction. Visual-to-audio errors are consistently larger than audio-to-visual errors under both SFT and FQFT, while FQFT reduces every reported directional rate.

Table 6: Directional cross-modal errors (%). V→A and A→V denote visual-to-audio and audio-to-visual interference. Lower is better.

Training	Detection Hall.		Quartet Hall.		Flip	
	V→A	A→V	V→A	A→V	V→A	A→V
SFT	30.00	22.30	15.57	2.77	7.61	0.55
+ FQFT	**14.40**	**9.80**	**11.45**	**2.10**	**3.89**	**0.00**

Positive-target recall. FQFT retains 93.23% recall on present audio targets and 92.18% on present visual targets in quartet evaluation, corresponding to omission rates of 6.77% and 7.82%. These recalls complement the false-presence rates and rule out uniformly absent predictions.

Analysis of training gains. FQFT jointly promotes target-label discrimination and score consistency across the four presence states: r_q separates present and absent label groups, n_q and i_q regularize other-label and interaction contrasts, and mean anchoring controls a shared score offset. The factorial representation makes these complementary constraints explicit, while the joint reward supervises answer-judgment correctness in complete responses. Because training quartets allow different clips under the same label, score consistency also constrains within-label margin variation. Auxiliary discrimination and within-label consistency are therefore constituent mechanisms of the quartet objective, rather than alternatives to its module-level contribution.

Table 3 crosses the presence or absence of the joint reward with the presence or absence of quartet regularization. All four policy variants share the structured-output SFT checkpoint, response schema, task-training data, answer/format rewards, and number of policy updates. Without the joint reward, adding Quartet increases JAA by 1.86 percentage points and reduces Hall. and Flip by 3.35 and 0.57 points. With the joint reward held fixed, adding Quartet increases JAA by 3.02 points and reduces Hall. and Flip by 4.26 and 0.33 points. Both comparisons thus support the incremental effectiveness of the quartet module under a fixed reward setting. Structured-output training is a shared foundation, not a condition unique to the Quartet variants.

The mathematical analysis supplies the design rationale for jointly constraining average and interaction contrasts: on matched inputs, $n_q = 0$ alone permits equal and opposite margin shifts when $i_q \neq 0$ (Equation 14). Constraining both components controls the two parallel changes; in the quadratic regime, their penalty is equivalently expressed through squared edge differences (Appendix D). This relation explains the organization of the consistency constraints, while the controlled ablation evaluates their combination with target discrimination and mean anchoring. The empirical attribution is to this combined module, rather than to the interaction term in isolation. The complete-response results measure its downstream effects through task-wise correctness and harmful transitions on matched inputs.

Empirical comparisons. The controlled ablation yields the highest joint correctness and lowest Hall. for FQFT, and the lowest Flip for quartet regularization alone. These scores are checkpoint-level point estimates under the stated training protocol; Flip describes conditional risk on each model's initially correct subset. The Qwen2.5-Omni ablations compare policy objectives from the same SFT initialization.

G Additional Cross-Benchmark Details

Evaluation scope. AVHBench* retains all source evaluation entries and media from AVHBench (Kim et al., 2025), adapting the task formulations to request explicit task-specific decisions together with the final answer. Table 4 compares SFT, SFT-plus-vanilla-GRPO, and SFT-plus-FQFT on these adapted external tasks. Relative to vanilla GRPO, FQFT increases AA, JA, and JAA by 5.64, 20.90, and 20.88 percentage points, respectively, and reduces Hall. by 3.17 points. Both policy-training variants share the same initialization obtained by supervised fine-tuning on 20K examples. The SFT row reports this shared checkpoint before policy optimization, using the same evaluation protocol as the policy-trained variants.

AVHBench contains 5,302 binary judgment questions and 1,106 captioning examples. Our adapted evaluation uses all 6,408 examples, grouped into 3,426 Joint examples, 1,876 AV Match examples, and 1,106 Caption examples. Correctness is averaged equally over Joint, AV Match, and Caption, whereas Hall. averages the two directional false-presence rates within Joint.

Correctness scoring. AA evaluates the final answer under the adapted response format. JA requires both target-presence decisions for Joint, the relation decision for AV Match, or the selected option content for Caption to be correct. JAA requires both the final answer and the required task-specific decisions to be correct. Caption is a four-choice description-selection task, not free-form caption generation; no generative judge is used for its reported scores.

For metric $m \in \{\mathrm{AA}, \mathrm{JA}, \mathrm{JAA}\}$, AVHBench* uses the equal-weight task average

$$m_{\mathrm{AVHBench}^*} = \frac{m_{\mathrm{Joint}} + m_{\mathrm{AVMatch}} + m_{\mathrm{Caption}}}{3}. \tag{16}$$

Directional hallucination scoring. Only the Joint task provides the absent-target conditions used for directional false-presence evaluation. The visual-to-audio direction contains 1,145 eligible absent audio targets, while the audio-to-visual direction contains 568 eligible absent visual targets.

Let $H_{V\to A}$ and $H_{A\to V}$ denote the corresponding false-presence rates. We report

$$\mathrm{Hall.}_{\mathrm{AVHBench}^*} = \frac{H_{V\to A} + H_{A\to V}}{2}. \tag{17}$$

Each directional rate uses its fixed gold-absent denominator.

Applicability of Flip. AVHBench* does not provide shared-clip quartets. Its examples therefore support directional false-presence measurement but not the controlled before-after comparison required by Flip, where the queried-modality input remains identical while only the other modality changes. We consequently report Hall. but not Flip on AVHBench*.

Prompt adaptation. The adapted prompts retain the source audio-visual examples. Joint and AV Match express the source judgments through explicit fields and answer options; Caption projects the source descriptions into a four-choice selection task. All evaluated systems receive the same

adapted questions and response schema through their respective input processors. The evaluated models receive the media and questions, but no gold labels or annotation metadata.

Human verification of the adapted protocol. The converted questions follow the same model-assisted pre-review and 2+1 human-review procedure as AV-TRACE (Appendix A). Two annotators independently inspect the original audio-visual media, task-specific reference labels, adapted options, and final answer mappings; disagreements are referred to a third annotator, with secondary review when needed to resolve ambiguity. For Joint, this includes checking both target-presence labels against their corresponding modalities. For Caption, the review checks whether the reference sentence is supported by the media and whether the alternative options preserve a unique correct selection. The complete original evaluation set is retained after prompt conversion and verification.

Media overlap. The 6,408 evaluation questions reference 2,136 distinct media files; an MP4 used for both audio and video input is counted only once. The SFT and reinforcement-learning subsets reference 24,000 and 52,845 unique training media files, respectively; their deduplicated union contains 67,287 files. We compare the media files used in training against the AVHBench evaluation media and find no identical files shared between the two sets. The comparison covers all media referenced by the 6,408 evaluation questions; the 191 additional files in the official media directory that are not referenced by these questions are excluded.

G.1 Task-to-field mapping and reference construction

Joint. Video-driven audio-hallucination questions use the source yes/no label for the audio target, with the visual target treated as present under the source task's visible-object condition. Audio-driven video-hallucination questions reverse this mapping: the source label determines visual presence, with the audio target treated as present under the audible-source condition. The accompanying-modality labels are initialized from these source task conditions (Kim et al., 2025); both modality labels are then checked against the original media through the human-review procedure described above. The two decisions determine the four-state label, which is mapped to the per-question final option. The combined-state field remains in the requested output, but the scored judgments are the audio and visual target decisions.

AV Match. The source yes/no matching label is mapped to matching or not_matching. The final option is determined by that question's two-option map. No new audio-target or visual-target presence labels are introduced for this task.

Caption. The adapted data contain four fixed sentence options per example and identify one reference option. The options and reference sentence are constructed from the source captioning examples and verified against the original media through the same human-review procedure. The original source caption is retained separately from the adapted reference sentence; they are not required to be verbatim identical. The response requests the selected sentence and its option letter. The parser maps the Selected content field to the supplied option contents using case and whitespace normalization. An explicit option-letter prefix can also identify the selected content when it matches its sentence or agrees with the separately reported final option, provided it does not contradict another exact option sentence. This is deterministic option-content scoring, not open-ended semantic judging or a measurement of free-form caption quality.

G.2 Complete adapted prompts and reference outputs

The examples below reproduce the archived prompts and gold reference responses. As in Appendix B.3, media markers indicate attached audio and video; the option text and mapping are instantiated for each example. They specify the adapted protocol used in Table 4, rather than the original AVHBench response format.

```
<video><audio>
Evaluate the sound event and visual entity named in the options. Use only the
  ↪ audio for the sound event and only the video for the visual entity; do not
  ↪ infer one modality from the other.

Which option is supported by the audio and video?

A. Neither target is supported.
B. Only the target visual "car" is supported.
C. The target sound "sound made by car" is audible and the target visual "car"
  ↪ is visible.
D. Only the target sound "sound made by car" is supported.

Return exactly the following structure. Do not add reasoning or extra sections.

[Decisions]
Audio target: <present or absent>
Visual target: <present or absent>
Joint state: <both_supported, audio_only, visual_only, or neither_supported>

[Final answer]
Answer: <A, B, C, or D>
```

Listing 4: AVHBench* Joint: complete adapted input prompt.

Joint: reference output.

```
[Decisions]
Audio target: present
Visual target: present
Joint state: both_supported

[Final answer]
Answer: C
```

```
<video><audio>
Judge the audio and video from their direct sensory evidence. Do not infer an
  ↪ unheard sound from a visible object or an unseen object from a sound.

Which option correctly describes their relationship?

A. The audio and video contexts are matching.
B. The audio and video contexts are not matching.

Return exactly the following structure. Do not add reasoning or extra sections.

[Decisions]
AV relation: <matching or not_matching>

[Final answer]
Answer: <A or B>
```

Listing 5: AVHBench* AV Match: complete adapted input prompt.

AV Match: reference output.

```
[Decisions]
AV relation: not_matching

[Final answer]
Answer: B
```

```
<video><audio>
Use only the directly observed audio and video evidence.

Which single-sentence description best matches what is seen and heard in the
  ↪ clip?

A. On a rainy street at night, a red-hooded man steals a parked black car while
  ↪ its owner chases him and shouts for help.
B. On a rainy street at night, a red-hooded driver narrowly avoids a pedestrian,
  ↪ remains inside the black car, and speaks to him through an open window.
C. On a dry street during the day, a driver stops beside a red-hooded man, gets
  ↪ out of a black car, and silently helps him cross the road.
D. On a rainy street at night, a driver hits a red-hooded man with a black car,
  ↪ gets out to check on him, and exchanges words with him as rain is heard.

Return exactly the following structure. Do not add reasoning or extra sections.

[Decisions]
Selected content: <copy the exact content of the selected option>

[Final answer]
Answer: <A, B, C, or D>
```

Listing 6: AVHBench* Caption: complete adapted input prompt.

Caption: reference output.

```
[Decisions]
Selected content: On a rainy street at night, a driver hits a red-hooded man
  ↪ with a black car, gets out to check on him, and exchanges words with him as
  ↪ rain is heard.

[Final answer]
Answer: D
```

H LIMITATIONS

AV-TRACE currently focuses on target-presence judgments and harmful absence-to-presence flips. Future work could extend the benchmark to temporal reasoning and open-ended descriptions, explore a broader range of cross-modal changes, and evaluate FQFT across more model families and training runs.