EGAA: EVIDENCE-GROUNDED ANSWER ALIGNMENT FOR MITIGATING CROSS-MODAL HALLUCINATION IN OMNI-MODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-modal hallucination remains a major challenge for omni-modal large language models. Existing benchmarks such as AVHBench and CMM measure these failures at the task level, but their standard scores provide limited information about the specific error within a generated response. We therefore examine generated answers at two complementary levels. Cross-Modal Source Misattribution (CSM), Absence Overgeneralization (AO), and Unsupported Certainty (UC) are identified at the statement level, while Wrong-Modality Answer (WMA) is assessed at the response level. To support both analysis and training, we augment AVHBench and CMM separately with visual and audio evidence annotations, supported statements, grounded answers, and error-targeted negative answers for these four error types. Based on these annotations, we propose Evidence-Grounded Answer Alignment (EGAA), a parameter-efficient post-training method that jointly learns from modality-specific evidence, supported statements, grounded answers, benchmark-format targets, and constructed preference pairs. We evaluate EGAA through bidirectional transfer between AVHBench and CMM. With Qwen2.5-Omni-7B, training on AVHBench improves CMM accuracy from 81.1% to 88.8%, while training on CMM improves AVHBench overall accuracy from 76.1% to 82.2%. These results show that explicitly connecting generated statements to multimodal evidence can reduce cross-modal hallucination across benchmarks. The code and configuration files used to reproduce our results are available at `https://anonymous.4open.science/r/EGAA-DF71/`, and the corresponding evidence annotations and dataset records are available at `https://osf.io/j7x89/overview?view_only=2ff822a2fee8471f9b6384c25d00ecf6`.

1 INTRODUCTION

Omni-modal language models integrate information from visual and audio streams to answer questions about their surroundings (Xu et al., 2025b; Cui et al., 2026; Zhang et al., 2023; Geng et al., 2025). This requires tracking which modality supports each claim (Zhang et al., 2026; Chung et al., 2026; Ye et al., 2025). AVHBench tests audio-driven video hallucination, video-driven audio hallucination, audio-visual matching, and captioning (Sung-Bin et al., 2025), while CMM studies failures associated with unimodal priors and spurious cross-modal correlations (Leng et al., 2025). Their task-level scores measure cross-modal reliability, but provide limited information about which claims in an explanatory answer are supported by each stream (Min et al., 2023; Jing et al., 2024; Rohrbach et al., 2018; Chen et al., 2024a).

We therefore analyze cross-modal hallucination at two complementary levels. By recording visual and audio evidence separately, we examine individual answer statements for CSM, AO, and UC, while WMA captures responses whose answer-bearing content addresses the wrong modality. These four recurring error types are illustrated in Figure 1.

This fine-grained view guides our data construction. We preserve the original media, questions, labels, task definitions, and official metrics of AVHBench and CMM, while adding separate visual and audio evidence annotations, supported statements, grounded answers, and error-targeted negative answers

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.

containing the four error types. The resulting annotations are used directly as the supervised targets and preference pairs for EGAA.

We evaluate whether the resulting supervision transfers beyond the benchmark from which it is constructed. Models trained with AVHBench supervision are evaluated on CMM under its official protocol, and models trained with CMM supervision are evaluated on AVHBench under its official protocol. With Qwen2.5-Omni-7B, EGAA improves CMM accuracy from 81.1% to 88.8% in the AVHBench→CMM direction and improves AVHBench overall accuracy from 76.1% to 82.2% in the CMM→AVHBench direction. We further evaluate EGAA on MiniCPM-o-4.5 and VideoLLaMA2.1-7B-AV using the same optimization recipe with backbone-specific multimodal input settings.

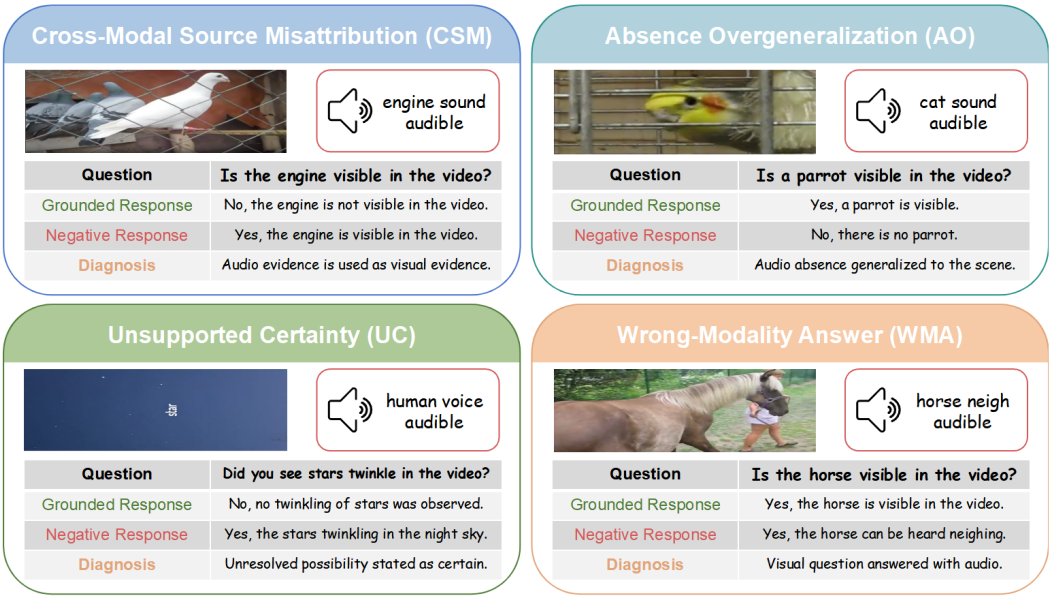


Figure 1: Four cross-modal error types used in this work.

Our contributions are threefold:

- **Fine-grained cross-modal error analysis.** We identify CSM, AO, UC, and WMA as recurring ways in which an answer can depart from the available multimodal evidence, providing a finer analysis than task-level accuracy alone.
- **Evidence-grounded annotations for existing benchmarks.** We augment AVHBench and CMM with visual and audio evidence, supported statements, grounded answers, and error-targeted negative answers, yielding 24,438 supported statements and 16,886 preference pairs while preserving the original tasks and official evaluation protocols.
- **Evidence-Grounded Answer Alignment.** We use the annotated evidence and supported statements to supervise answer generation, and the four error types to construct closely matched preference pairs. The resulting joint supervised and preference training is evaluated through bidirectional cross-dataset transfer.

2 Related Work

2.1 Audio-Visual Hallucination Benchmarks

AVHBench (Sung-Bin et al., 2025) was designed to test whether an audio-visual model keeps the source of evidence straight when answering a binary question. Its audio-driven video and video-driven audio tasks ask about one modality while the other modality can create a misleading association. The matching task asks whether the audio and visual contexts correspond, and the captioning task asks for a joint description. This mixture is useful because a model can fail by inventing a visual object from an audio cue, by inventing a sound from a visible object, or by judging two individually

plausible streams to be related when they are not (Wang et al., 2024). CMM (Leng et al., 2025) takes a complementary route: its 2,400 probes are organized around spurious inter-modality correlations and over-reliance on a unimodal prior, with visual, audio, and language dominance conditions. The benchmark pairs positive and negative probes so that perception and resistance to hallucination can be read separately. These designs make the two benchmarks good tests of cross-modal reliability, but their official scores still reduce a generated response to a task-level decision.

We retain these benchmark tasks and labels and add annotations that connect individual statements to the evidence available in each modality. This distinction matters for transfer: a correct short answer may still be accompanied by an explanation that attributes evidence to the wrong modality (Sung-Bin et al., 2025; Leng et al., 2025; Jung et al., 2025b). The statement-level annotations allow us to examine these explanations alongside the official task scores.

2.2 Fine-Grained Factual Evaluation

FActScore (Min et al., 2023) established a useful evaluation pattern for long-form generation: split an answer into atomic factual units and evaluate support for those units instead of assigning one label to the whole answer. FaithScore (Jing et al., 2024) adapts the same idea to vision-language generation. It first isolates descriptive portions of a response, extracts atomic facts, and checks each fact against the image (Chen et al., 2024b). The shared insight is that a response can contain both useful and unsupported content, and that the unsupported unit identifies the failure more precisely than a response-level score (Manakul et al., 2023; Zha et al., 2023).

Audio-visual answers require two additional distinctions. A claim can be true but attributed to the wrong modality, and a negative observation can be restricted to the supplied audio or frames without implying that the event is absent from the world (Sung-Bin et al., 2025; Leng et al., 2025). Conversely, a relation between streams, such as whether an engine sound belongs to the visible aircraft, is an inference from both sides rather than an observed concept in either side alone (Sung-Bin et al., 2025; Wu et al., 2024). We therefore retain atomic statements while recording the source, scope, requested modality, and certainty of the claim. These annotations serve both as a diagnostic reference and as targets for constructing error-targeted preference answers.

2.3 Training-Time Alignment for Cross-Modal Hallucination

RLHF-V (Yu et al., 2024) shows how fine-grained corrections can be used directly in dense preference optimization: human annotators correct the response segment by segment, and the preferred response is trained against the corrected alternative (Sun et al., 2024; Yu et al., 2025). MoD-DPO (Chaubey et al., 2026) addresses the audio-visual case by making the preference objective (Rafailov et al., 2023) depend on the modality that the question requires. It keeps the answer stable when an irrelevant stream is changed, increases sensitivity to changes in the relevant stream, and suppresses text-only shortcuts. The two lines of work differ in supervision, but both make the rejected answer informative about the failure being corrected (Jiang et al., 2024; Sarkar et al., 2025).

EGAA pairs a grounded answer with an error-targeted negative answer generated from modality-specific evidence. Separate supervised targets train the model on the evidence trace, grounded answer, and benchmark-format answer, while the preference objective distinguishes grounded answers from these error-targeted negative answers. The cross-dataset experiments evaluate how well this supervision transfers to a different benchmark.

3 Method

Our method connects evidence annotation, answer construction, and model training. We first define three statement-level errors involving the source, scope, or certainty of a claim, together with a response-level error for answering about the wrong modality. We then annotate each benchmark item in three stages: separate visual and audio evidence, a fused description with supported statements and a grounded answer, and error-targeted negative answers. Finally, the same annotations supervise three response formats and define the error-targeted negative answers for preference training. At test time, the model receives only the supplied media, question and output-format instruction; the textual evidence is used to construct training targets and to audit the generated explanation. Figure 1

illustrates the errors, Figure 2 shows the annotation stages, and Figure 3 shows how the targets are jointly optimized and evaluated.

3.1 Cross-Modal Error Types

We derive the error types from responses generated by the base Qwen2.5-Omni-7B (Xu et al., 2025a) model on AVHBench. Using the reviewed visual and audio evidence, descriptions, supported statements, and grounded answers described in Section 3.2 as references, we collect erroneous responses and provide the full questions, responses, and annotations to Qwen3.7-Plus to organize recurring patterns. The authors review the resulting groups and define four error types according to the relation among the requested modality, the modality-specific evidence, and the claim expressed in the response. Each atomic statement expresses one proposition and is evaluated together with the full question and response context.

Each target concept receives one of four modality-specific statuses: `observed`, `possible`, `not_observed`, or `unavailable`. The first three characterize the evidence within a provided modality, whereas `unavailable` indicates that the modality is not provided for that item, as can occur in CMM. The visual and audio evidence are annotated independently, while their combination can further support conclusions represented by the supported statements and grounded answer. We use these evidence states and the question to characterize four cross-modal error types, illustrated in Figure 1:

- **Cross-Modal Source Misattribution (CSM):** A concept is `observed` or `possible` in one modality and `not_observed` in the other, but the response claims that the concept is directly observed in the latter.
- **Absence Overgeneralization (AO):** A concept is `observed` in at least one modality, but the response generalizes a modality-specific non-observation into a claim that the concept is absent from the entire scene or event.
- **Unsupported Certainty (UC):** The evidence does not establish a cross-modal source for an `observed` event or resolve an interpretation marked `possible`, but the response states the source relation or interpretation as certain.
- **Wrong-Modality Answer (WMA):** For a question about one modality, the answer-bearing content addresses only the other modality and gives no conclusion about the requested one.

These four error types describe complementary relations among the question, evidence, and response: CSM concerns the modality assigned to a concept, AO concerns the scene-level scope of an absence claim, UC concerns certainty that exceeds the available evidence, including unsupported event–object associations and unresolved possible interpretations, and WMA concerns the modality addressed by the answer. CSM, AO, and UC are determined by the status of each modality and the relationship between the two streams, whereas WMA is determined by the question and the answer-relevant content.

These evidence relations also determine how the response is annotated. When the two streams jointly support the same event source, this shared source is reflected in the conclusion. The resulting evidence annotations are then used to construct error-targeted negative answers for preference training. Labels are assigned at the level of atomic statements: each statement receives at most one error label, although different statements within the same response may exhibit different error types.

3.2 Evidence-Grounded Data Construction

We augment AVHBench and CMM separately with the annotations used for training and fine-grained evaluation. Each benchmark item retains its original media, question, raw answer, and task definition, and receives modality-specific evidence, a description, supported statements, a grounded answer, and error-targeted negative answers. The pipeline keeps the two benchmarks separate so that the resulting annotations can be used for bidirectional transfer without changing either benchmark's official evaluation protocol.

For an audio-visual input $x = (v, a, q)$, we parse the question to identify its intent, relevant concepts, and requested modalities. A sample may contain several visual and audio concepts. We annotate visual

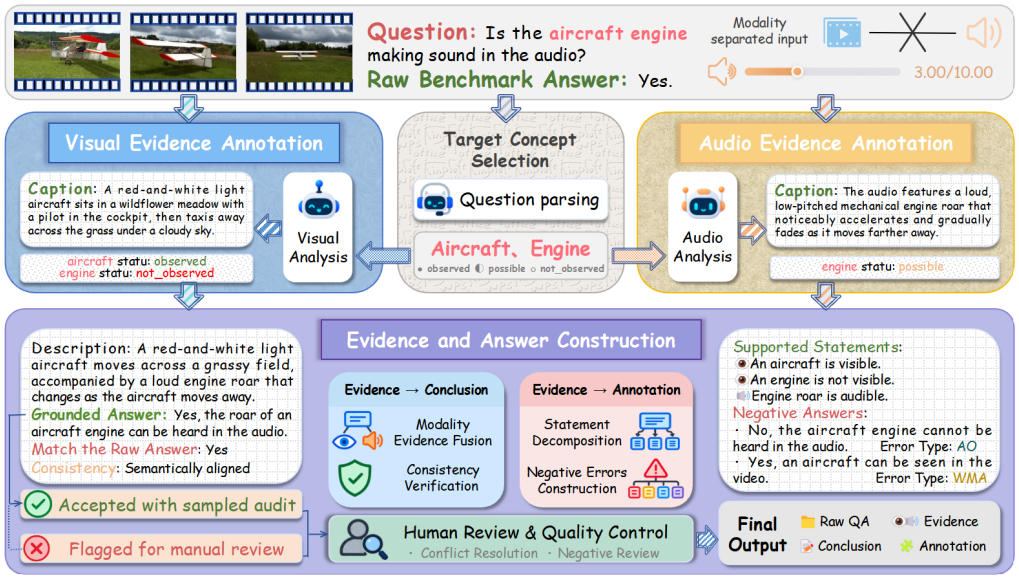


Figure 2: Construction pipeline for the evidence-grounded annotations.

evidence E_V from sampled frames and audio evidence E_A from the audio stream independently. Each concept receives a description and a modality-specific status: `observed` for identifiable evidence, `not_observed` when the concept is not observed in the supplied modality, or `possible` when the available evidence supports a plausible but unresolved interpretation. These statuses record direct observations within each modality. CMM inputs retain only the modalities actually provided; an unavailable modality is described as `unavailable`, rather than assigned `not_observed`.

The fusion model combines the visual and audio annotations with the original question to produce a description, supported statements, and a grounded answer. Joint evidence can resolve an ambiguity left by one modality. In Figure 2, for example, the aircraft's motion and the changing engine sound support an attribution of the sound to the aircraft, while the audio-only status remains `possible`. The aircraft is visible, but its engine is not directly visible. The modality-specific statuses record the observations, and the supported statements express conclusions drawn from their combination. The raw benchmark answer is used only after generation for consistency checking; it is not provided to the fusion model. We compare the generated answer with the raw benchmark answer, send disagreements and ambiguities to human review, and conduct sampled audits of accepted annotations.

Table 1: Corpus and preference-supervision statistics for AVHBench and CMM.

	Statistic	AVHBench	CMM	Total
Corpus	Benchmark items	6,408	2,400	8,808
	Supported statements	19,191	5,247	24,438
	Negative answers	19,291	6,208	25,499
Preference	CSM pairs	3,786	452	4,238
	AO pairs	3,640	771	4,411
	UC pairs	2,056	324	2,380
	WMA pairs	2,761	3,096	5,857
	All preference pairs	**12,243**	**4,643**	**16,886**

Matching questions use the same evidence pipeline but ask about a relation between the streams. For an `AV_MATCHING_QUERY`, we first keep the visual and audio concept annotations separate. The fusion model then compares their content and records whether the contexts match, do not match, or provide insufficient evidence. This relation is an inference, so it is not written as `observed`, `not_observed`, or `possible` for either modality. Resolved judgments determine the grounded Yes/No answer and are checked against the raw benchmark answer; insufficient-evidence cases enter

human review. Captioning items use the same separated evidence but express the supported content as a description rather than a matching relation.

Finally, we construct error-targeted negative answers from the modality-specific evidence and supported statements. CSM places a concept on the modality side where it is not supported, AO turns observed content into a claim of global absence, UC creates an unsupported certain event–object association, and WMA redirects a single-modality answer to the other stream. The construction prompt receives the question, evidence annotations, supported statements, and a designated error type, and returns a candidate for that type. We randomly select at most two error-targeted negative answers per item and pair them with the grounded answer for preference training.

Annotation construction covers all 6,408 AVHBench items—5,302 judgment questions and 1,106 captioning items—and all 2,400 CMM items (Sung-Bin et al., 2025; Leng et al., 2025). The resulting corpus sizes and preference-pair counts are summarized in Table 1.

3.3 Evidence-Grounded Answer Alignment

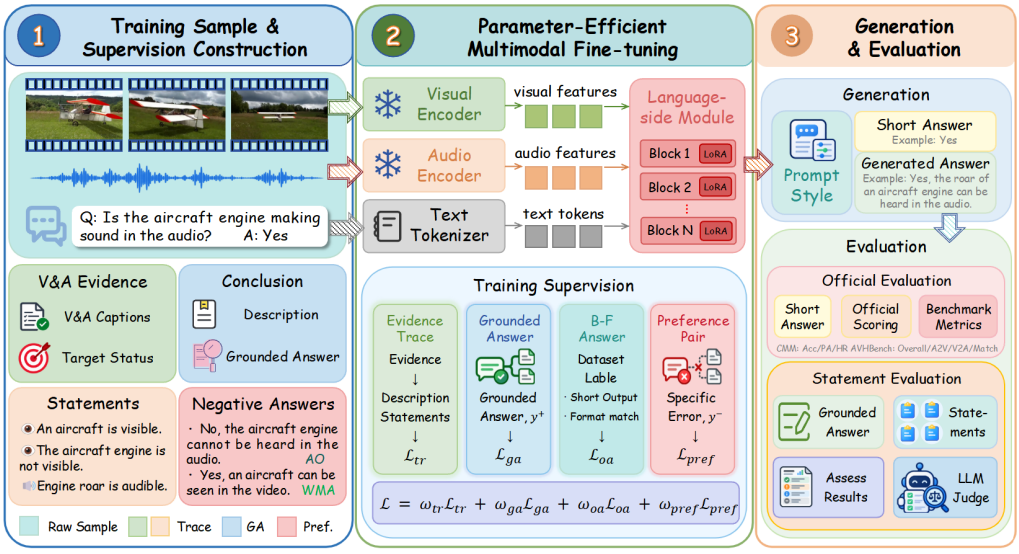


Figure 3: Overview of the EGAA training and evaluation framework.

The annotations constructed in Section 3.2 provide the training signals for EGAA. As shown in Figure 3, EGAA applies LoRA (Hu et al., 2022) to the language-side module while keeping the visual and audio encoders frozen.

We use four training signals: an Evidence Trace (Trace), a Grounded Answer, a Benchmark-Format Answer, and a Preference Pair (Pref.). The Evidence Trace concatenates the modality-specific evidence, combined description, and supported statements, followed by the Benchmark-Format Answer. The Grounded Answer is the reviewed explanatory response derived from the combined evidence. The Benchmark-Format Answer is the training target in the benchmark's output format: a reviewed short answer for judgment questions and the descriptive target for captioning items. A Preference Pair contrasts the grounded answer with an error-targeted negative answer constructed in Section 3.2.

The three supervised targets share the same media and question and use instructions that specify the requested response format. Trace requests separate visual and audio evidence, a description, supported statements, and a final `[Answer]` section containing the Benchmark-Format Answer. Grounded Answer uses the original question and the grounded answer as its target. Benchmark-Format Answer appends "Answer only Yes/No" for judgment questions and uses the reviewed short answer; for captioning items included in the training data, the benchmark descriptive target is retained. Let p_k and y_k denote the instruction and target for $k \in \{tr, ga, bf\}$. Each supervised loss is defined as

$$\mathcal{L}_k = -\mathbb{E}_{(x,p_k,y_k)} \log \pi_\theta(y_k \mid x, p_k). \tag{1}$$

The resulting losses are $\mathcal{L}_{tr}$, $\mathcal{L}_{ga}$, and $\mathcal{L}_{bf}$. Appendix A.2 gives the prompt formats.

For preference optimization, each pair contains a grounded answer y^+ and an error-targeted negative answer y_t^- that instantiates one of the four error types. Using the corresponding base model as the reference policy, we optimize the direct preference objective of DPO (Rafailov et al., 2023):

$$\mathcal{L}_{\text{pref}} = -\mathbb{E} \log \sigma \left(\beta \left[\log \frac{\pi_\theta(y^+ \mid x)}{\pi_{\text{ref}}(y^+ \mid x)} - \log \frac{\pi_\theta(y_t^- \mid x)}{\pi_{\text{ref}}(y_t^- \mid x)} \right] \right). \tag{2}$$

Here, β controls the strength of the preference objective.

The three supervised objectives and the preference objective are optimized jointly, with all four losses updating the same LoRA parameters in the language-side module:

$$\mathcal{L} = w_{tr}\mathcal{L}_{tr} + w_{ga}\mathcal{L}_{ga} + w_{bf}\mathcal{L}_{bf} + w_{pref}\mathcal{L}_{pref}. \tag{3}$$

The w terms control the relative contributions of the four training objectives.

At inference time, the model receives the available media, question, and an instruction specifying the requested output format. Official judgment evaluation uses short answers. Fine-grained error evaluation uses a separate fixed prompt requesting a concise explanatory answer. The judge evaluates the generated answer against the evidence annotations, supported statements, and grounded answer. These references are provided to the judge, while the evaluated model receives only the media, question, and evaluation prompt.

4 Experiments and Results

4.1 Setup

We evaluate bidirectional cross-dataset transfer. In AVHBench→CMM, training uses the 6,408 augmented AVHBench items and evaluation uses the 2,400 CMM questions. In CMM→AVHBench, training uses the 2,400 augmented CMM items and evaluation uses the 5,302 AVHBench judgment questions. Each model is trained on the source benchmark and evaluated on the other benchmark.

For dataset construction, Qwen3.7-Plus annotates visual evidence and combines the visual and audio annotations into descriptions, supported statements, and grounded answers. Qwen3.5-Omni-Plus extracts audio evidence. The resulting annotations supply the training targets described in Section 3.2.

We evaluate Qwen2.5-Omni-7B, MiniCPM-o-4.5 (Cui et al., 2026), and VideoLLaMA2.1-7B-AV (Cheng et al., 2024). For each backbone, we compare Base, Benchmark-Format only, EGAA, and component ablations. Benchmark-Format only learns the benchmark-format targets. The ablations remove the Trace, Grounded Answer, or Preference loss from the joint objective. Appendix A records the implementation settings.

CMM reports accuracy, perception accuracy (PA) on positive cases, and hallucination resistance (HR) on negative cases. AVHBench reports Overall, Audio→Visual (A2V), Visual→Audio (V2A), and Matching accuracies. To examine the content behind these short judgments, we also prompt Qwen2.5-Omni-7B to answer each question with Yes or No followed by a brief explanation. The same instruction is used for Base, Benchmark-Format only, and EGAA.

DeepSeek-V4-Flash evaluates these explanations against the reviewed evidence annotations, supported statements, and grounded answer. It assesses WMA at the response level and checks atomic statements for CSM, AO, and UC. Miss records an atomic statement that preserves the modality, scope, and evidential commitment of the relevant observation but misidentifies or incorrectly describes directly observable content. It is reported separately from the four cross-modal hallucination categories. Table 4 counts each question once in every applicable category; Any error is the union of all five categories. Table 5 reports the proportion of extracted statements supported by the evidence. Both analyses use the same 2,400 CMM and 5,302 AVHBench judgment questions. Appendix A.4 details the labeling and aggregation.

4.2 RESULTS

In Table 2, EGAA improves CMM Accuracy and AVHBench Overall over the corresponding base model across all three backbones. These gains in both transfer directions indicate that the evidence-grounded supervision remains useful beyond its source benchmark. Related methods and their reported results are summarized in Appendix B.1; because their training data, optimization, or decoding settings differ from ours, they are kept separate from this controlled comparison.

Table 2: Official cross-dataset judgment results and component ablations (%).

		CMM			AVHBench			
Backbone	Method	Acc	PA	HR	Overall	A2V	V2A	Match
Qwen2.5-Omni-7B	Base	81.1	92.7	69.6	76.1	79.0	75.4	75.1
	Benchmark-Format only	86.1	89.2	83.0	72.7	82.9	84.0	52.7
	EGAA w/o Trace	88.0	85.8	90.2	81.0	84.2	84.5	74.7
	EGAA w/o GA.	88.4	89.2	87.6	81.8	84.9	82.0	79.8
	EGAA w/o Pref.	87.6	82.4	92.8	80.7	85.4	83.7	74.1
	EGAA	**88.8**	88.2	89.3	**82.2**	84.3	82.1	81.0
MiniCPM-o-4.5	Base	85.4	90.3	80.6	79.6	84.7	78.4	78.0
	Benchmark-Format only	87.1	92.8	81.3	72.5	87.1	74.6	61.0
	EGAA w/o Trace	87.9	88.0	87.8	80.2	84.9	80.0	77.6
	EGAA w/o GA.	88.2	89.5	86.9	81.5	86.0	82.5	77.5
	EGAA w/o Pref.	87.5	84.5	90.5	79.9	85.6	78.5	78.1
	EGAA	**88.6**	89.1	88.2	**82.3**	86.9	86.7	74.0
VideoLLaMA2.1-7B-AV	Base	82.8	89.2	76.4	67.5	76.9	70.5	58.1
	Benchmark-Format only	83.5	90.0	77.0	71.1	79.1	72.8	64.2
	EGAA w/o Trace	84.3	85.5	83.1	85.8	87.0	87.3	83.2
	EGAA w/o GA.	84.7	86.3	83.1	86.9	86.6	87.0	86.9
	EGAA w/o Pref.	83.9	82.8	85.0	85.3	85.1	85.6	85.0
	EGAA	**85.0**	85.8	84.3	**87.8**	87.9	88.1	87.3

On CMM, EGAA substantially improves hallucination resistance while keeping perception accuracy close to the Base level. The models reject unsupported claims more reliably while retaining most of their ability to recognize positive cases. The preference ablation helps explain this pattern. Without preference supervision, HR is highest for all three backbones, but PA is also lowest. Adding preference supervision improves positive-case recognition while keeping HR well above Base, yielding higher overall accuracy. This suggests that the grounded and error-targeted negative answers help the model distinguish supported claims from unsupported ones while preserving sensitivity to positive evidence.

Benchmark-Format only also improves CMM Accuracy for all three backbones, but transfers less consistently to AVHBench. Qwen2.5-Omni-7B and MiniCPM-o-4.5 gain in A2V yet lose substantial Matching accuracy, lowering their Overall scores. This contrast suggests that short-label supervision learned on CMM transfers more readily to questions about one modality than to judgments of correspondence between streams. VideoLLaMA2.1-7B-AV improves on all three AVHBench tasks. Its lower initial Matching accuracy leaves more room for improvement, which may partly explain its different response to fine-tuning. EGAA improves Matching over Benchmark-Format only for every backbone. MiniCPM-o-4.5 still falls below Base on Matching, although its stronger A2V and V2A results raise Overall. Thus, the aggregate gains reflect different patterns of improvement across backbones and tasks.

Table 3: CMM challenge-wise accuracy across backbones (%).

Backbone	Method	Spurious Corr.	Unimodal Priors	Visual-dom.	Audio-dom.	Language-dom.	Event	Object
Qwen2.5-Omni-7B	Base	89.8	72.4	69.8	67.8	**79.8**	80.3	82.3
	Benchmark-Format only	91.0	81.3	78.0	88.8	77.0	83.2	90.2
	EGAA	**91.6**	**85.9**	**89.0**	**89.3**	79.5	**87.3**	**90.8**
MiniCPM-o-4.5	Base	87.9	82.9	85.8	83.3	79.8	83.1	88.7
	Benchmark-Format only	89.3	84.8	83.8	89.8	**81.0**	84.6	90.6
	EGAA	**91.8**	**85.4**	**86.8**	**91.5**	78.0	**86.7**	**91.3**
VideoLLaMA2.1-7B-AV	Base	89.9	75.7	72.8	83.5	**70.8**	80.6	85.8
	Benchmark-Format only	90.0	77.0	75.5	86.0	69.5	81.4	86.4
	EGAA	**90.3**	**79.8**	**81.8**	**90.5**	67.0	**82.7**	**88.3**

The supervised components also contribute to transfer. Removing either Trace or Grounded Answer lowers CMM Accuracy and AVHBench Overall for every backbone. Trace links the answer to explicit

observations and supported statements, while Grounded Answer directly supervises an evidence-based response to the question. The full objective performs best when these targets are learned together with preference supervision.

The challenge breakdown in Table 3 shows gains over Base in six of the seven reported CMM categories for every backbone. Qwen2.5-Omni-7B improves most on visual- and audio-dominant probes, which directly test the tendency to substitute evidence from one stream for another. These gains are consistent with the separation of visual and audio observations during training. Language-dominant accuracy, however, declines across backbones. This category concerns reliance on linguistic priors over visual evidence, whereas EGAA focuses on the source and support of audio-visual claims. The difference in emphasis offers a plausible explanation for the less favorable result in this category, alongside improvements on the other cross-modal challenges.

Table 4: Question-level error incidence for Qwen2.5-Omni-7B (%).

Target	Method	CSM	AO	UC	WMA	Miss	Any error
CMM	Base	5.00	7.21	11.75	0.38	0.33	21.79
	Benchmark-Format only	1.79	1.58	10.29	0.38	0.33	14.21
	EGAA	1.54	1.21	9.21	0.04	0.25	12.17
AVHBench	Base	2.38	6.21	10.77	1.15	4.60	24.31
	Benchmark-Format only	3.39	3.26	16.52	1.32	8.58	31.86
	EGAA	2.36	3.23	10.75	0.21	4.04	20.24

The explanatory-answer analysis in Table 4 further distinguishes the effects of label supervision and evidence grounding. On AVHBench, Benchmark-Format only reduces AO but increases UC and Miss, leaving more questions with at least one error. EGAA retains the reduction in AO, brings UC close to Base, and lowers WMA and Miss. Relative to Base, its clearest reductions are in AO and WMA, with little change in CSM and UC. On CMM, Benchmark-Format only already reduces CSM, AO, and UC; EGAA reduces all four hallucination categories further and also lowers Miss. The lower combined error incidence on both benchmarks is therefore accompanied by different changes in the constituent error types.

Table 5: Statement-level evidence support for Qwen2.5-Omni-7B explanatory answers.

Target	Method	Statements	Support (%)
CMM	Base	3,611	82.1
	EGAA	3,607	85.6
AVHBench	Base	8,087	65.9
	EGAA	8,047	73.8

Table 5 examines how much of each explanatory answer is grounded in the evidence. EGAA increases the proportion of supported statements on both benchmarks, with a larger gain on AVHBench. The number of extracted statements remains nearly unchanged, suggesting that the improvement comes from better-supported content at a comparable level of detail. This complements the question-level error analysis, which counts a question once in each applicable category regardless of how many statements contain that error. Statement-level support also reflects changes within answers that may still contain an error.

5 Conclusion

We introduced four cross-modal error types and used them to construct evidence-grounded annotations and preference pairs. EGAA jointly learns from evidence traces, grounded answers, benchmark-format targets, and these preferences. It improves aggregate judgment accuracy in both transfer directions between AVHBench and CMM across three backbones. For Qwen2.5-Omni-7B, the explanatory answers contain fewer questions with errors and a higher proportion of supported statements at comparable statement counts. These results support connecting answer supervision to modality-specific evidence to improve cross-modal reliability.

AI USE STATEMENT

Generative AI tools were used in data construction to annotate modality-specific evidence and construct descriptions, supported statements, grounded answers, and error-targeted negative answers. They were also used for language editing and for organizing the data and code for open-source release. The authors determined the final error definitions, reviewed the constructed annotations, and take responsibility for the final content and released materials.

REPRODUCIBILITY STATEMENT

Section 3 specifies the evidence annotations, supervised targets, and joint training objective. Section 4 describes the evaluation tasks and metrics. Appendix A records optimization and evaluation details, and Appendix B specifies the cross-dataset transfer protocol. We release the source code, annotation files, and implementation details through the anonymous repository linked in the abstract to facilitate independent verification and reproduction of our experiments.

REFERENCES

Ashutosh Chaubey, Jiacheng Pang, and Mohammad Soleymani. Mod-dpo: Towards mitigating cross-modal hallucinations in omni llms using modality decoupled preference optimization. *ArXiv*, abs/2603.03192, 2026. URL https://api.semanticscholar.org/CorpusID:286229091.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Chao Sun, Rongzhou Zhang, Guanyu Zhou, and Lijie Wen. Omnidpo: a preference optimization framework to address omni-modal hallucination. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i24.39104. URL https://doi.org/10.1609/aaai.v40i24.39104.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. Unified hallucination detection for multimodal large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3235–3252, 2024a.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Jianing Yang, David Fouhey, Joyce Chai, and Shengyi Qian. Multi-object hallucination in vision language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 44393–44418. Curran Associates, Inc., 2024b. doi: 10.52202/079017-1409. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4ea4a1ea4d9ff273688c8e92bd087112-Paper-Conference.pdf.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. URL https://arxiv.org/abs/2406.07476.

Sangyun Chung, Se Yeon Kim, Youngchae Chee, and Yong Man 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, June 2026.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18959–18969, 2025. doi: 10.1109/CVPR52734.2025.01766.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27026–27036, 2024. doi: 10.1109/CVPR52733.2024.02553.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. FaithScore: Fine-grained evaluations of hallucinations in large vision-language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5042–5063, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.290. URL https://aclanthology.org/2024.findings-emnlp.290/.

Chaeyoung Jung, Youngjoon Jang, Jongmin Choi, and Joon Son Chung. Fork-merge decoding: Enhancing multimodal understanding in audio-visual large language models. *arXiv preprint arXiv:2505.20873*, 2025a.

Chaeyoung Jung, Youngjoon Jang, and Joon Son Chung. Avcd: Mitigating hallucinations in audio-visual large language models through contrastive decoding. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 63143–63174. Curran Associates, Inc., 2025b. doi: 10.52202/085713-2121. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/5b5e70e7fb3a9060562361b2385346df-Paper-Conference.pdf.

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. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-3601. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/9b0b18a77421d45d26c3df5612caefe7-Paper-Datasets_and_Benchmarks_Track.pdf.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Suzanne Petryk, David M. Chan, Anish Kachinthaya, Haodi Zou, John Canny, Joseph E. Gonzalez, and Trevor Darrell. ALOHa: A new measure for hallucination in captioning models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 342–357, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.30. URL https://aclanthology.org/2024.naacl-short.30/.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model Is Secretly a Reward Model. In

ICML 2023 Workshops: MFPL, 2023. URL https://mlanthology.org/icmlw/2023/rafailov2023icmlw-direct/.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437/.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Arik, and Tomas Pfister. Mitigating object hallucination in mllms via data-augmented phrase-level alignment. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 73796–73846, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/b73228d026ef49bfa49f95b7e330513e-Paper-Conference.pdf.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the association for computational linguistics: ACL 2024*, pp. 13088–13110, 2024.

Kim Sung-Bin, Oh Hyun-Bin, Lee Jung-Mok, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 24244–24271, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/3cc685788a311fa35d8d41df93e288ca-Paper-Conference.pdf.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024.

Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. Evaluating and analyzing relationship hallucinations in large vision-language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53553–53570. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wu24l.html.

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.

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.

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

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-Grained Correctional Human Feedback. In *Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024. doi: 10.1109/CVPR52733.2024.01310. URL https://mlanthology.org/cvpr/2024/yu2024cvpr-rlhfv/.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, et al. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19985–19995. IEEE, 2025.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. AlignScore: Evaluating factual consistency with a unified alignment function. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11328–11348, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.634. URL https://aclanthology.org/2023.acl-long.634/.

Chenshuang Zhang, Kyeong Seon Kim, Chengxin Liu, and Tae-Hyun Oh. Svhalluc: Benchmarking speech–vision hallucination in audio-visual large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2026.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In Yansong Feng and Els Lefever (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 543–553, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.49. URL https://aclanthology.org/2023.emnlp-demo.49/.

A Implementation and Reproducibility Details

A.1 Optimization and Decoding

Each cross-dataset transfer model is fine-tuned for one epoch. The joint loss weights are $(w_{tr}, w_{ga}, w_{bf}, w_{pref}) = (1, 1, 0.5, 1)$. Qwen2.5-Omni-7B uses greedy decoding for evaluation.

For Qwen2.5-Omni-7B, LoRA uses rank 16, $\alpha = 32$, dropout 0.05, and the language-side linear projections in the transformer layers; visual and audio encoders remain frozen. The adapted projections are the attention q, k, v, and o projections together with the MLP gate, up, and down projections. We use AdamW with learning rate 1×10^{-4}, zero weight decay, a warmup ratio of 0.03, and gradient clipping at 1.0. Training uses bfloat16, gradient checkpointing, micro-batch size 1, gradient accumulation of 16, and a maximum sequence length of 6,144 tokens. The random seed is 17. DPO uses $\beta = 0.1$ and at most two preference pairs per item. Official judgment decoding uses greedy decoding with at most 16 new tokens; the explanatory prompt allows at most 256 new tokens.

For Qwen2.5-Omni-7B, evaluation samples video at 2.0 frames per second, with 4–768 frames and at most 602,112 pixels per frame. Training uses 1.0 frame per second, at most 48 frames, and at most 100,352 pixels per frame. Audio is resampled to 16 kHz and truncated to 30 seconds. MiniCPM-o-4.5 and VideoLLaMA2.1-7B-AV use 16 uniformly sampled frames for evaluation.

MiniCPM-o-4.5 and VideoLLaMA2.1-7B-AV use the same optimization recipe as Qwen2.5-Omni-7B: one training epoch; $(w_{tr}, w_{ga}, w_{bf}, w_{pref}) = (1, 1, 0.5, 1)$; LoRA rank 16, $\alpha = 32$, and dropout 0.05; AdamW with learning rate 1×10^{-4}, zero weight decay, warmup ratio 0.03, and gradient clipping at 1.0; bfloat16; gradient checkpointing; micro-batch size 1; gradient accumulation 16; maximum sequence length 6,144; and seed 17. DPO uses $\beta = 0.1$ and at most two preference pairs per item. Visual and audio encoders remain frozen, and LoRA is applied to the language-side attention q, k, v, and o projections and the MLP gate, up, and down projections. Official judgment uses greedy decoding with at most 16 new tokens, while explanatory prompts allow at most 256 new tokens.

Evaluation uses 16 uniformly sampled frames for MiniCPM-o-4.5 and VideoLLaMA2.1-7B-AV. Audio follows each backbone's native preprocessor; when a separate waveform is required, it is resampled to 16 kHz and truncated to 30 seconds. MiniCPM-o-4.5 training caps each clip at 8 frames with `max_slice_nums=1`, while evaluation retains 16 frames. VideoLLaMA2.1-7B-AV uses 16 frames for both training and evaluation, consistent with the public checkpoint's `num_frames` setting.

The experiments use Qwen2.5-Omni-7B, MiniCPM-o-4.5, and VideoLLaMA2.1-7B-AV. Training runs on two NVIDIA RTX A6000 GPUs with distributed data parallelism. The per-device micro-batch size is 1 and gradient accumulation is 16, giving a global effective batch size of 32. The Qwen experiments use Python 3.12.13, PyTorch 2.6.0 with CUDA 12.4, and Transformers 4.52.3; VideoLLaMA2.1-7B-AV training uses PyTorch 2.2.0 with CUDA 12.1. We use one seed (17) for each backbone and transfer direction, and reuse the same recipe for the ablations.

Table 6 reports the wall-clock time for the main EGAA runs.

Table 6: Wall-clock training time for the main EGAA runs on two RTX A6000 GPUs.

Backbone	AVHBench→CMM	CMM→AVHBench
Qwen2.5-Omni-7B	9.1 h	4.5 h
MiniCPM-o-4.5	8.2 h	2.9 h
VideoLLaMA2.1-7B-AV	16.1 h	4.6 h

The ablation switches correspond to the losses in the main objective. Benchmark-Format only enables $\mathcal{L}_{bf}$; w/o Trace removes $\mathcal{L}_{tr}$; w/o GA. removes $\mathcal{L}_{ga}$; and w/o Pref. removes $\mathcal{L}_{pref}$. The released configuration files record these switches.

For joint EGAA training, Yes/No Benchmark-Format Answer targets are retained with probability 0.5 using seed 17. AVHBench captioning items always retain their descriptive Benchmark-Format Answer. The Benchmark-Format only ablation retains all eligible Yes/No targets. Preference sampling uses the same seed and at most two negative answers per item.

A.2 Prompt Templates and Output Formats

The training and evaluation pipeline uses fixed prompt templates. Question parsing identifies the inspection targets and requested modalities without access to the benchmark answer. Visual and audio evidence are then annotated independently from their respective streams. Evidence fusion receives only the question and the structured modality-specific annotations. Negative-answer construction uses the question, modality-specific evidence, supported statements, and one designated error type.

Question parsing. You are a strict question parser. Inspect only the question; do not answer it and do not use or infer the benchmark label. Identify what the question asks, the concepts that must be inspected, and the modality or modalities required to answer it. A concept is an inspection target only and must not be treated as evidence that the concept is actually present. Distinguish visual targets, audio targets, and relations between the two streams according to the wording of the question. For an audio-visual matching question, identify that the question asks about the relation between the visual and audio contexts rather than the presence of either concept alone. For a captioning request, identify that both available modalities may contribute to the requested description. Use concise lowercase English snake_case for concept names. Return one JSON object containing the question intent, the requested modalities, and the concepts to inspect. Do not add evidence, answers, confidence values, or Markdown outside the JSON.

Visual evidence. You are a strict visual-evidence analyst. Inspect only the uniformly sampled frames from the video; you cannot hear the audio. Report objects, actions, attributes, and scene properties that are directly visible. Do not use common sense, the question wording, or an audio cue to add a fact that is not visible. For every prompted concept, provide a short evidence description and exactly one status: observed when it is clearly visible, not_observed when it does not appear in the supplied frames, possible when the available evidence supports a plausible but unresolved interpretation, and unavailable when the visual modality is missing for the item. Prompted concepts are inspection targets and do not imply that they occur in the frames. If a prompt names a sound concept such as dog_bark or engine_sound, check the corresponding visible entity when one exists, for example dog or engine/car, and state the visual limitation in the evidence. Also add salient visible objects, actions, or scene properties that were not prompted, marking them observed. Use lowercase English snake_case for concept names. Return one JSON object with the keys `description` and `evidence`; each evidence item has `concept`, `status`, and `evidence`. Do not add quality scores, confidence values, or Markdown outside the JSON.

Audio evidence. You are a strict audio-evidence analyst. Inspect only the audio extracted from the video; you cannot see the frames. Report sound events, speech, music, and environmental sounds that are directly audible. Do not use common sense, the question wording, or a visual cue to add a sound that is not audible. For every prompted sound concept, provide a short evidence description and exactly one status: observed when it is clearly audible, not_observed when it is not

heard in the supplied audio, possible when the available evidence supports a plausible but unresolved interpretation, and unavailable when the audio modality is missing for the item. Hearing a sound establishes that the sound is present, but does not establish that its source is visible in the video. If a prompt names a visual concept such as dog or car, check the corresponding sound event when one exists, for example dog_bark or engine_sound, and state the audio limitation in the evidence. Also add salient audible events that were not prompted, marking them observed. Use lowercase English snake_case and make the concept name denote a sound event, such as dog_bark, engine_sound, rain_sound, male_speech, background_music, or footsteps. Return one JSON object with the keys `description` and `evidence`; each evidence item has `concept`, `status`, and `evidence`. Do not add quality scores, confidence values, or Markdown outside the JSON.

Evidence fusion. You are a strict multimodal evidence-fusion analyst. You receive only the question and the structured visual and audio evidence; you cannot inspect the original media and you must not use or infer the official benchmark label. Use only information supported by the supplied evidence. Preserve every modality-specific concept and status exactly as given: do not change observed, not_observed, possible, or unavailable based on the other stream. Produce a concise joint description that summarizes the relevant evidence without inventing unobserved details. Then produce atomic supported statements, with each statement expressing one proposition that is directly supported by one modality or justified by combining the two streams. Joint evidence may resolve an ambiguity that remains possible within one modality, but such a conclusion must appear only as a supported statement and must not replace or rewrite the original modality-specific status. Do not treat not_observed in one modality as absence from the entire scene or event, and do not treat an audible source as visually observed or a visible source as audibly observed unless the corresponding modality provides that evidence. For AV_MATCHING_QUERY, keep the visual and audio concepts separate and include exactly one relation statement stating that the contexts match, do not match, or that the evidence is insufficient. If a modality required by the question is unavailable, state that it is unavailable for judgment rather than treating it as not_observed. Finally, produce a grounded answer that directly answers the question using only the supported evidence and conclusions. For a resolved Yes/No question, begin the grounded answer with Yes or No and give a concise evidence-based explanation; do not force a resolved judgment when the required evidence is unavailable or insufficient. Return one JSON object with the keys `description`, `statements`, and `grounded_answer`. Do not add confidence scores, quality ratings, or Markdown outside the JSON.

Negative-answer construction. You are a strict negative-answer constructor. You receive the question, modality-specific visual and audio evidence, supported statements, and one requested error type: CSM, AO, UC, or WMA. Construct one negative answer from the annotated evidence for the requested type. Preserve the relevant evidence-supported content and express exactly one targeted error. The candidate must instantiate exactly the requested error and must otherwise preserve the meaning, relevant details, and level of certainty of the grounded answer. Do not introduce a second error or unrelated unsupported content. For CSM, use a concept marked `observed` or `possible` in one modality and `not_observed` in the other; make the answer place that concept on the latter side. For AO, use a concept marked `observed` in at least one modality and make the answer claim that it is absent from the whole scene or event; two `not_observed` statuses do not provide an AO candidate. For UC, state a cross-modal source relation involving an observed event or a possible interpretation as certain when the evidence does not establish or resolve it. For WMA, use a single-modality question and answer only about the other modality, with either polarity and without a conclusion about the requested modality. Apply WMA to modality-specific questions rather than descriptions or audio–visual relations. Before returning the candidate, check it against the question, the modality-specific evidence, the supported statements, and the requested error type. Reject candidates that do not express the requested error or that contain an additional error. Return only the requested error type and the single negative answer in a JSON object; do not add explanations, scores, or Markdown outside the JSON.

Official and supervised targets. For a judgment item, the Benchmark-Format Answer uses the instruction "Answer only Yes/No" and the reviewed short answer used for training. The raw benchmark answer is retained separately for official evaluation and consistency checking. For a captioning item, the Benchmark-Format Answer uses the benchmark descriptive target. The Trace instruction asks the model to list the visual and audio evidence separately with their modality-specific statuses, followed

by a concise combined description, supported statements, and a final answer; its target ends with an `[Answer]` section containing the Benchmark-Format Answer. The Grounded Answer contains only the reviewed evidence-grounded answer to the original question. These three supervised targets use the same available media and question while specifying different output formats.

Fine-grained error judging. You are a strict fine-grained evaluator of a generated multimodal answer. You receive the question, the modality-specific visual and audio evidence, the annotated supported statements, the grounded answer, and the generated response. Treat the supplied evidence and reference annotations as the basis for evaluation; do not add facts from common sense or from information not present in these inputs. First determine whether the answer-bearing content addresses the modality requested by the question. Mark WMA when the response instead answers about another modality. Then extract the explanatory content of the generated response into atomic statements, with each statement expressing one proposition. Do not create explanatory statements from a bare Yes/No answer. Evaluate each extracted statement in the full question and response context. Assign at most one statement-level error label according to the type of proposition being distorted. CSM applies when a concept is observed or possible in one modality and not observed in the other, but the response claims that the concept is directly observed in the latter. AO applies when a concept is observed in at least one modality, but the response generalizes a modality-specific non-observation into a claim that the concept is absent from the entire scene or event. UC applies when the evidence does not establish a cross-modal source for an observed event or resolve an interpretation marked possible, but the response states the source relation or interpretation as certain. Miss applies when a statement preserves its modality, scope, and evidential commitment but misidentifies or incorrectly describes directly observable content. If a response contains multiple propositions, split them into atomic statements before labeling. Each erroneous atomic statement receives one of CSM, AO, UC, or Miss according to whether the error concerns source attribution, absence scope, evidential commitment, or directly observable content. Separately assess evidence support for every extracted statement: mark it supported only when its source, scope, content, and certainty are justified by the modality-specific evidence or the annotated supported statements. A statement can be factually related to the scene but still be unsupported if it assigns the evidence to the wrong modality, broadens its scope, or increases its certainty beyond the evidence. Return one structured JSON object containing the response-level WMA decision and the extracted atomic statements with their error labels and support decisions. Do not add quality ratings, confidence scores, or Markdown outside the JSON.

A.3 Annotation provenance and consistency

Each record contains modality-specific evidence, a description, supported statements, a grounded answer, and constructed negative answers. The raw benchmark answer is retained for post-generation consistency checks. The audit sample and its resolution counts are reported below.

Two annotators audited a pooled sample of 800 AVHBench items and 300 CMM items during dataset construction. The pool combines label–annotation conflict cases with stratified random draws of non-conflict items. Among the 760 AVHBench disagreement cases, 686 followed the annotation: the modality evidence and grounded answer were retained, and only the short Yes/No field used as the Benchmark-Format Answer was corrected; 74 followed the raw benchmark answer, so the evidence, supported statements, and grounded answer were rewritten. Among the 281 CMM disagreement cases, 182 followed the annotation and 99 followed the raw benchmark answer. The remaining 40 AVHBench and 19 CMM items were non-conflict spot checks, all of which passed.

All official accuracy metrics are computed against the raw benchmark labels.

The raw benchmark answer remains in the original dataset label. The released annotation files store the original benchmark label in `raw_answer` and the reviewed short answer in `benchmark_format_answer`. The reviewed explanatory answer is stored in `grounded_answer`, supported statements in `statements`, and error-targeted negative answers in `negative_answers`. At most two valid negative answers are sampled per item for preference training. Miss is used only in statement-level evaluation and is not a preference-training category.

Separately, we audit the automated fine-grained judgments used at evaluation time. DeepSeek-V4-Flash provides the statement-level judgments used in our fine-grained evaluation. Two annotators manually reviewed 200 AVHBench responses (340 extracted statements) and 100 CMM responses

(123 extracted statements), covering all three backbone models and both transfer directions. For each sampled response, the annotators checked the response-level WMA decision and the labels assigned to its extracted statements, with the human annotations treated as the reference.

No human-annotated error was classified as correct by the evaluator. On AVHBench, eight human-labeled MISS statements were classified as UC, and four were classified as AO; additionally, one response judged correct by the annotators was classified as WMA. On CMM, five human-labeled MISS statements were classified as UC, while two statements judged correct by the annotators were classified as AO. We also identified one incomplete statement that was nevertheless assigned a MISS label.

A.4 FINE-GRAINED ERROR EVALUATION DETAILS

The evaluator receives the question, modality-specific evidence, annotated supported statements, grounded answer, and generated response. WMA is assessed at the response level for questions about one modality. CSM, AO, and UC are checked for each extracted statement according to the definitions in Section 3.1. Miss records atomic statements that misidentify or incorrectly describe directly observable content while preserving its modality, scope, and evidential commitment. Each question contributes at most once to each category, and Any error counts the union of the five categories.

Evidence support is assessed separately for every extracted statement. Support is the number of supported statements divided by the total number of extracted statements, pooled across answers for each dataset and method. Statement-level generation uses greedy decoding, and at most 256 new tokens. DeepSeek-V4-Flash judges each response with temperature 0.0, a maximum of 2,048 output tokens, and thinking disabled. It first assesses WMA at the response level, then extracts atomic statements and labels statement-level errors as CSM, AO, UC, or Miss according to the definitions above (Petryk et al., 2024). Statement extraction is performed by the judge rather than by a fixed regular expression. Bare Yes/No-only outputs contribute no explanatory statements, and empty or unparseable generations are excluded from support pooling. Question-level rates count each category at most once, and Any error is the union of the five categories.

B CROSS-DATASET TRANSFER PROTOCOL

The transfer protocol uses the complete augmented source benchmark for training and the other benchmark for evaluation. The two benchmarks contain no overlapping media.

The released Qwen2.5-Omni-7B workflow includes corpus construction, training, inference, official scoring, and ablation configurations for both transfer directions.

The code is released for research reproducibility. Benchmark media and Qwen2.5-Omni weights remain subject to their original licenses. The released annotation files do not redistribute the original media.

B.1 COMPARISON WITH RECENT CROSS-MODAL HALLUCINATION METHODS

We additionally compare EGAA with recent methods that report results on AVHBench or CMM using the same or closely related audio-visual backbones. Because the published studies use different training corpora, optimization recipes, decoding procedures, model versions, and evaluation implementations, the following results are provided as contextual references rather than controlled head-to-head comparisons. We preserve the values reported in the original studies and include only metrics that are directly comparable to those used in our evaluation.

MoD-DPO (Chaubey et al., 2026) augments DPO with modality-sensitivity and modality-invariance regularization, encouraging the model to remain invariant to corruption of an irrelevant modality while remaining sensitive to corruption of the modality required by the question. MoD-DPO++ further introduces a language-prior debiasing term. The original study reports results for Qwen2.5-Omni-7B and MiniCPM-o-2.6 on both AVHBench and CMM. In contrast, EGAA constructs preference supervision from modality-specific evidence, supported statements, grounded answers, and error-targeted negative answers.

Table 7: Reported MoD-DPO results and our EGAA results on AVHBench and CMM (%).

		AVHBench			CMM	
Backbone	Method	A2V Acc.	V2A Acc.	Match Acc.	PA	HR
Qwen2.5-Omni-7B	Base	84.15	77.38	54.69	86.4	84.6
	DPO	84.39	79.68	59.32	85.2	84.6
	OmniDPO	85.34	80.77	61.50	86.6	84.7
	MoD-DPO	87.66	82.48	69.07	88.8	86.2
	MoD-DPO++	88.19	83.40	69.68	89.2	87.2
	EGAA	84.3	82.1	81.0	88.2	89.3
MiniCPM-o-2.6	Base	83.36	74.54	54.26	85.6	80.4
	DPO	82.91	78.86	54.56	85.2	80.3
	OmniDPO	84.96	75.39	56.86	86.4	80.6
	MoD-DPO	87.08	79.00	60.57	88.0	82.5
	MoD-DPO++	87.26	79.49	60.66	88.3	83.6
MiniCPM-o-4.5	EGAA	86.9	86.7	74.0	89.1	88.2

Fork-Merge Decoding (FMD) (Jung et al., 2025a) is a training-free inference strategy for reducing modality bias in audio-visual language models. It processes audio-only and video-only representations separately in the early decoder layers and merges the resulting hidden states for joint reasoning in later layers. The original study evaluates FMD on AVHBench with VideoLLaMA2, video-SALMONN, and Qwen2.5-Omni, reporting classification accuracy for the A2V, V2A, and AV Matching tasks. These three metrics directly correspond to the AVHBench judgment metrics used in our evaluation. Table 8 reports the results for the two backbones that are the same as or closely related to those evaluated with EGAA.

Table 8: Reported FMD and EGAA results on AVHBench (%).

			AVHBench		
Setting	Backbone	Method	A2V Acc.	V2A Acc.	Match Acc.
Reported	Qwen2.5-Omni	Vanilla	80.77	71.20	77.45
		FMD	81.30	71.77	78.22
	VideoLLaMA2	Vanilla	80.02	77.03	57.75
		FMD	80.45	77.52	59.01
Ours	Qwen2.5-Omni-7B	Base	79.0	75.4	75.1
		EGAA	84.3	82.1	81.0
	VideoLLaMA2.1-7B-AV	Base	76.9	70.5	58.1
		EGAA	87.9	88.1	87.3

OmniDPO (Chen et al., 2026) constructs text-preference and multimodal-preference pairs to strengthen audio–visual interaction and reduce reliance on language priors. Its preference construction differs from EGAA in that it does not use statement-level evidence traces or negative answers targeted to the four cross-modal error types. The original study evaluates Qwen2.5-Omni-7B and MiniCPM-o-2.6 on AVHBench and CMM. We report the A2V and V2A accuracies from AVHBench together with the PA and HR metrics from CMM.

Table 9: Reported OmniDPO results and our EGAA results on AVHBench and CMM (%).

		AVHBench		CMM	
Backbone	Method	A2V Acc.	V2A Acc.	PA	HR
Qwen2.5-Omni-7B	Base	74.12	67.60	91.7	70.2
	VCD	77.40	68.11	92.0	71.3
	ICT	76.70	68.80	92.1	71.8
	DPO	71.74	68.70	92.8	73.3
	OmniDPO	84.42	77.51	93.3	74.7
	EGAA	84.3	82.1	88.2	89.3
MiniCPM-o-2.6	Base	74.65	72.80	88.1	63.9
	VCD	73.97	73.96	88.5	63.8
	ICT	75.92	72.25	88.9	65.1
	DPO	76.07	74.73	90.1	66.2
	OmniDPO	76.85	76.68	91.0	68.8
MiniCPM-o-4.5	EGAA	86.9	86.7	89.1	88.2