Don't Let the Video Speak: Audio-Contrastive Preference Optimization for Audio-Visual Language Models

Ami Baid, Zihui Xue, Kristen Grauman

University of Texas at Austin

Abstract. While Audio-Visual Language Models (AVLMs) have achieved remarkable progress over recent years, their reliability is bottlenecked by cross-modal hallucination. A particularly pervasive manifestation is video-driven audio hallucination: models routinely exploit visual shortcuts to hallucinate expected sounds, discarding true auditory evidence. To counteract this deeply ingrained visual dominance, we propose Audio-Contrastive Preference Optimization (ACPO). This dual-axis preference learning framework introduces an output-contrastive objective to penalize visual descriptions masquerading as audio facts, alongside an input-contrastive objective that swaps audio tracks to explicitly penalize generation invariant to the true auditory signal. Extensive experiments demonstrate that ACPO establishes highly faithful audio grounding and mitigates audio hallucination without compromising overarching multimodal capabilities.

Keywords: audio-visual language models · cross-modal hallucination · video question-answering

1 Introduction

Large language models (LLMs) [1,5,46] have transformed natural language processing, serving as flexible interfaces for summarization [25,41], question answering [21,26], and reasoning [22,49]. But language alone is insufficient for understanding the physical world; human perception seamlessly integrates language, vision, and sound. Vision-language models (VLMs) extend LLMs by incorporating visual input [3,28,30]. Audio-visual language models (AVLMs) emerged following VLMs, able to jointly reason over text, video, and audio [10,44,52].

Audio provides critical context that vision alone cannot capture: it alerts us to off-screen events, clarifies the hidden mechanics of object interactions, and conveys the nuanced emotions of speakers. Integrating both auditory and visual streams enables AVLMs to drive meaningful advancements in real-world applications, like autonomous systems [40] and assistive technologies [2,18] that must reliably interpret their surroundings. Through generative tasks like question answering and captioning, these models provide a flexible, user-friendly interface to translate complex multimodal environments into accessible insights.

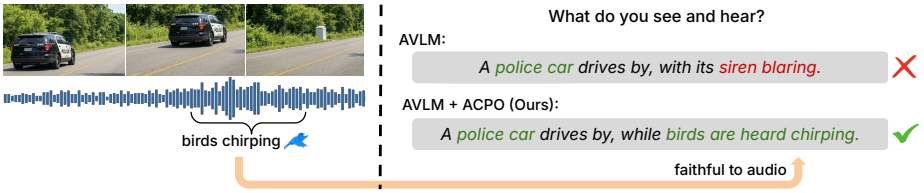

Fig. 1: Correcting cross-modal hallucination in AVLMs. Current models often exploit real-world co-occurrence shortcuts, leading them to hallucinate sounds based on what is seen rather than what is heard. Here, the base AVLM incorrectly predicts a siren due to the visual presence of a police car. Our approach explicitly corrects this class of modality attribution errors, decoupling the visual shortcut to accurately ground the response in the actual audio track.

Despite their impressive fluency, LLMs struggle with hallucination, generating text that is plausible but factually ungrounded [17, 19]. When extended to the visual domain, VLMs inherit this vulnerability and exhibit visual hallucination—confidently describing entities, attributes, or actions that are entirely absent from the visual input even if semantically plausible [15, 29, 38]. Consequently, as models evolve to process increasingly diverse sensory streams, maintaining strict factual grounding across multiple modalities emerges as a central challenge.

The integration of audio and video in AVLMs introduces a uniquely challenging variant of this problem: *cross-modal hallucination*. Unlike unimodal errors, these hallucinations are driven by misleading priors across sensory streams, where a model generates claims that are plausible under one modality but unsupported by the full context. Relying on visual priors, for example, a model might hallucinate an accompanying siren for a silent police car (see Fig. 1); conversely, auditory cues might trigger descriptions of off-screen speakers as if they were visible. These errors are deeply insidious; they mimic natural audio-visual co-occurrences and read with perfect fluency. Yet, they expose a critical flaw in the model's modality attribution, undermining the reliability of AVLM generation.

Crucially, cross-modal hallucination is an asymmetric phenomenon. Existing literature demonstrates that AVLMs systematically default to visual priors. Audio tokens receive disproportionately low attention weights during decoding [20], and models are notably more prone to hallucinating on audio-focused tasks than visual ones [23]. We observe this exact asymmetry in practice (Sec. 3.1): feeding more video frames into AVLMs actively triggers more audio hallucinations, while adding audio causes no such degradation to visual QA tasks. This dynamic exposes a deep-seated visual dominance: models are overwhelmingly prone to guessing sounds based on what they see. Addressing this vulnerability is the core focus of our work, as we aim to enforce true auditory grounding and ensure models *don't let the video speak.*

Despite its pronounced impact on model trustworthiness, video-driven audio hallucination is an under-explored vulnerability. Early attempts to address

this—spanning both training-free [20] and training-based [7] methods—have proven insufficient, as they either do not penalize cross-modal leakage or leave the weak audio representations unimproved. Deeply investigating this capability gap, we identify two compounding factors at the root of visual dominance. (1) The strong correlation between sight and sound in large-scale datasets [6, 31] encourages models to rely on superficial co-occurrence shortcuts, bypassing the need to learn true modality attribution. (2) The inherent architectural imbalance, where vision encoders benefit from vastly larger datasets and stronger supervision than audio encoders [20, 23, 43], heavily biases the model. As a direct result, visual priors dictate the response, and conflicting audio evidence is systematically ignored [20, 23].

To fundamentally correct this visual dominance, we propose Audio-Contrastive Preference Optimization (ACPO). Our key insight is that naively applying preference learning to static audio-visual inputs fails to break co-occurrence shortcuts, as models can still guess the correct audio response using solely visual cues. Instead, ACPO forces true modality attribution by constructing preference pairs along two orthogonal axes. First, we employ output-contrastive pairs using audio-swapped inputs, explicitly penalizing the model for generating visually-driven descriptions in response to audio queries. Second, we introduce input-contrastive pairs that evaluate identical text outputs across different audio tracks, penalizing the model if its predictions remain invariant when the supporting auditory evidence is removed. By jointly optimizing these pairs and fine-tuning only the audio projection layer, our lightweight framework strengthens audio grounding without disrupting the backbone's established vision-language capabilities.

Extensive evaluations validate the efficacy of our framework. To rigorously isolate and measure modality-specific grounding in free-form generation, we introduce a novel unimodal captioning evaluation that probes audio and visual understanding under both aligned and mismatched conditions. Across this rigorous new protocol and established standard hallucination benchmarks, ACPO elevates isolated audio captioning quality and decisively curtails video-driven audio hallucinations, while preserving general multimodal capabilities. Ultimately, our findings establish that targeted, contrastive modality supervision is a highly effective and necessary mechanism for achieving balanced, trustworthy response generation in AVLMs.

2 Related Work

2.1 Audio-visual Language Models

AVLMs extend vision-language models (VLMs) by incorporating audio as an additional input stream [42, 44, 54]. In a typical architecture, a video encoder [36, 47] and an audio encoder [8] produce modality-specific embeddings, which are projected into a shared space and fed alongside tokenized text into an LLM backbone. Early AVLMs [42,54] demonstrated the viability of this paradigm, and more recent models [10, 44, 52] achieve competitive performance on benchmarks requiring joint audio-visual reasoning [27, 31, 53].

2.2 Hallucination in LLMs and VLMs

Hallucination has been studied extensively in LLMs [17,19] and VLMs [15,29,38]. Training free strategies include chain-of-thought prompting [49], self consistency sampling [13], and contrastive decoding methods [24], which adjust outputs logits to reduce reliance on language priors. Training-based approaches include hallucination-aware supervised objectives [16,50], auxiliary alignment losses [12], and preference based learning [37, 51]. DPO [37] has emerged as a particularly effective framework, able to steer model outputs without catastrophic forgetting [55]. Unlike policy-gradient methods such as PPO [39], DPO operates on fixed preference pairs without requiring an explicit reward model or on-policy rollouts. V-DPO [51] applies DPO to visual hallucination by constructing preference pairs using out-of-distribution objects (e.g., cutting a rock instead of a cake). These methods are effective for vision-language grounding, but they address only a single non-text modality and do not consider interactions between multiple input streams.

2.3 AVQA Benchmarks

Cross-modal hallucination in AVLMs remains relatively underexplored. Nishimura et al. [32] offered early analysis of audio hallucination in AVLMs, finding that models frequently generate audio descriptions driven by visual content rather than the audio signal. Subsequent benchmarks probe this failure more systematically. AVHBench [43] evaluates cross-modal hallucination using questions whose ground truth response depends on a single modality while the other may be misleading. CMM [23] decomposes hallucination into unimodal dominance and spurious inter-modality correlations, measuring both under controlled masking and corruption. Both benchmarks confirm that current AVLMs are especially vulnerable when audio and video conflict.

To address this, audio-visual contrastive decoding [20] masks selected modalities at inference time and adjusts logits to reduce hallucination, but incurs additional cost and does not modify underlying representations. OmniDPO [7] extends preference optimization to the audio-visual setting. For each example, they add noise to the audio or video stream and train the model to prefer responses conditioned on the clean input over the corrupted one. However, noisy inputs remain broadly aligned with the original video, and therefore do not directly address video-driven audio hallucination. Our approach differs by constructing preference pairs with cross-modal conflict that elicit hallucination, and by introducing modality-specific supervision that penalizes visually hallucinated audio descriptions.

3 Method

We introduce Audio-Contrastive Preference Optimization (ACPO), a lightweight preference-based framework that strengthens audio grounding in AVLMs. We describe the setup and background in Sec. 3.1, and detail ACPO in Sec. 3.2.

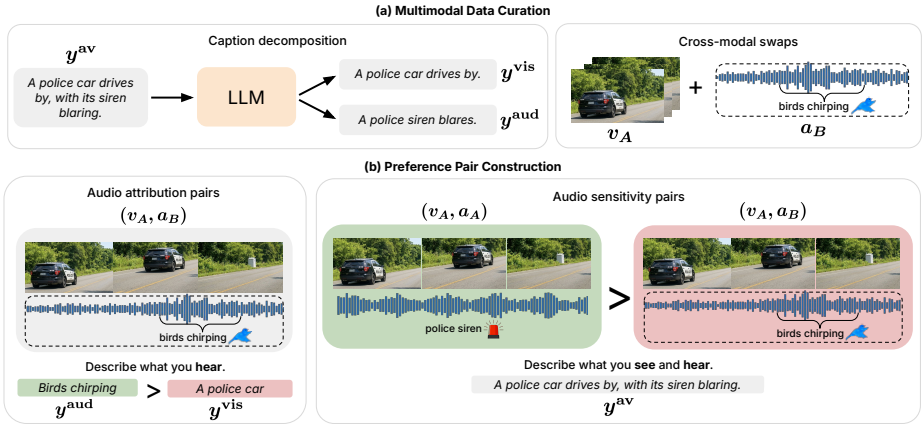

Fig. 2: Overview of ACPO. (a) Multimodal Data curation: each joint audio-visual caption is decomposed into modality-specific targets y^{vis} and y^{aud}, and audio-swapped inputs (v_A, a_B) are constructed by replacing the original audio track with a mismatched one. (b) Preference pair construction: audio-attribution pairs (left) use the swapped input (v_A, a_B) to penalize visually-driven responses to audio-focused prompts, preferring y^{aud} over y^{vis}. Audio-sensitivity pairs (right) penalize audio-invariant predictions by preferring the original audio-visual caption y^{av} under the aligned input (v_A, a_A) over the same caption under the swapped input (v_A, a_B).

3.1 Problem Setup

We consider an AVLM that receives a video v, an audio track a, and a text prompt x, and generates a text response y. The model $p_\theta(y \mid v, a, x)$ encodes v and a through modality-specific encoders, projects the resulting representations into a shared embedding space via learned projection modules, and feeds the projected tokens alongside the tokenized prompt into an LLM backbone. Our objective is to ensure that the model intelligently integrates cues from both v and a, maintaining factual grounding across both modalities rather than hallucinating content from one single modality.

Empirically, we identify a pronounced asymmetry in modality reliance, where models disproportionately favor visual cues. This imbalance reflects a capability gap, as vision encoders and visual training pipelines are typically stronger and more mature than their audio counterparts [20, 23, 43]. As shown in our analysis (Fig. 3, left), when evaluating audio-focused question answering, progressively adding visual information (by scaling the number of input frames) actively degrades performance. This suggests that rather than providing complementary context, the added visual input confuses the model and overrides valid audio evidence. Conversely, augmenting video-focused QA tasks with the audio modality does not result in a similar performance drop (right), indicating that the model is relatively robust to audio interference but highly susceptible to visual dominance.

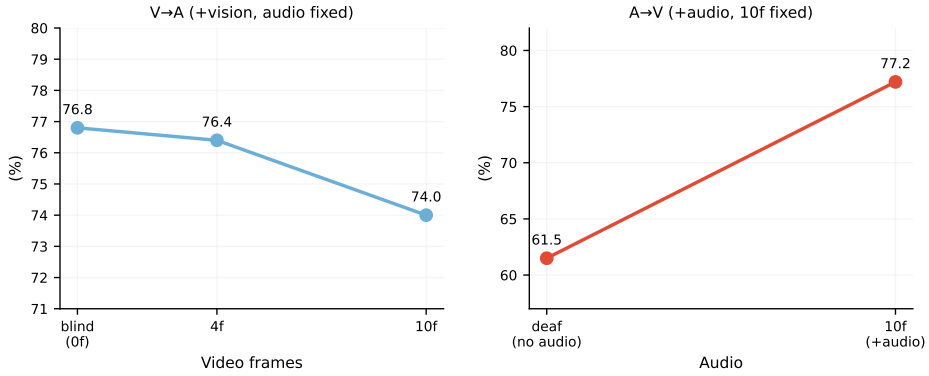

Fig. 3: Visual dominance in AVLMs is asymmetric. On the AVHBench [43] video-driven audio hallucination task (V→A), adding more video frames progressively degrades performance (76.8 → 74.0), as visual priors override auditory evidence. On the audio-driven video hallucination task (A→V), adding audio improves performance (61.5 → 77.2), indicating the model is robust to audio interference but highly susceptible to visual dominance.

Motivated by this vulnerability to visual interference, we introduce Audio-Contrastive Preference Optimization (ACPO). ACPO is designed to counteract visual dominance, ensuring that the model remains faithfully grounded in the actual audio content instead of hallucinating sounds driven by strong visual priors.

3.2 Audio-Contrastive Preference Optimization

Correcting visual dominance requires teaching the model to trust the audio signal even when the visual signal provides a tempting, but potentially misleading, prior. We formulate this as a preference learning problem. Our foundation is Direct Preference Optimization (DPO) [37], which aligns language models by contrasting a preferred response y^+ against a dispreferred response y^- conditioned on an input context c:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma\Big(\beta\big[\log \tfrac{p_\theta(y^+|c)}{p_{\text{ref}}(y^+|c)} - \log \tfrac{p_\theta(y^-|c)}{p_{\text{ref}}(y^-|c)}\big]\Big), \tag{1}$$

where p_{ref} is the frozen reference model and β controls the deviation penalty. Intuitively, DPO increases the relative likelihood of the preferred response while anchoring updates to the reference distribution.

Our key insight is that naively applying standard DPO—which relies on a fixed, joint audio-visual context c—fails to isolate the root cause of video-driven hallucination. Because real-world audio and video are highly correlated, a model can learn to output the "preferred" audio caption merely by exploiting visual co-occurrence shortcuts, defeating the purpose of the alignment. To force true audio grounding, ACPO instead constructs preference pairs along two axes: one

contrasts *what the model says* (output-contrastive), the other contrasts *what the model hears* (input-contrastive). The pair types rely on modality-specific supervision targets and controlled audio–visual mismatches, which we describe first. Fig. 2 illustrates both the data curation pipeline and the resulting preference pairs.

Caption decomposition. Standard audio-visual captions entangle visual and auditory evidence in a single sentence, offering no signal for modality-specific grounding. Starting from standard audio-visual captions, we use a large language model to decompose each joint caption into a visual caption y^{vis} describing only visually verifiable content and an audio caption y^{aud} describing only audibly verifiable content. These decomposed captions serve as modality-specific supervision targets for the preference pairs below.

Audio-swapped inputs. To decouple audio from its co-occurring visual context, we construct audio-swapped inputs that retain the original video while replacing the audio track. For a source clip with video v_A and audio a_A, we sample a partner clip and form the swapped input (v_A, a_B). To control the degree of audio-visual mismatch, we measure the similarity between the original video and candidate audio tracks using a multimodal embedding model [14], grouping swaps into high- and low-similarity tiers based on empirical quantiles, where low-similarity swaps introduce stronger audio-visual conflict.

Audio-attribution pairs (output-contrastive). These pairs penalize the model for describing visual content in response to an audio-specific prompt. For a clip with video v_A and swapped track a_B, we prompt the model with an audio-focused instruction x_{aud} (e.g., "Describe what you hear."). The preferred response is the audio caption y_B^{aud} corresponding to the actual audio track a_B; the dispreferred response is the visual caption y_A^{vis} for the video v_A. The visual caption may be fluent and plausible given the video, but it reflects visual rather than auditory evidence. These pairs directly penalize the model for sourcing its response from the wrong modality, enforcing the preference:

$$\log p_\theta(y_B^{\text{aud}} \mid v_A, a_B, x_{\text{aud}}) > \log p_\theta(y_A^{\text{vis}} \mid v_A, a_B, x_{\text{aud}}). \tag{2}$$

Audio-sensitivity pairs (input-contrastive). These pairs penalize the model when its predictions are invariant across different audio tracks, indicating that it is not using auditory evidence. For a fixed video v_A, we consider both the aligned input (v_A, a_A) and an audio-swapped input (v_A, a_B). Let y_A^{av} denote the original audio-visual caption for clip A, and let x be a joint instruction (e.g., "Describe what you see and hear."). The preferred configuration is $((v_A, a_A),\ y_A^{\text{av}})$; the dispreferred configuration is $((v_A, a_B),\ y_A^{\text{av}})$. Since a_B does not support the audio-specific content in y_A^{av}, the model should assign lower likelihood to the same caption when the supporting auditory evidence has been replaced. These pairs penalize audio-insensitive behavior, enforcing the preference:

$$\log p_\theta(y_A^{\text{av}} \mid v_A, a_A, x) > \log p_\theta(y_A^{\text{av}} \mid v_A, a_B, x). \tag{3}$$

During training, we optimize the DPO objective (Eq. 1) over the union of all pair types, fine-tuning only the audio projection layer to improve audio grounding without disrupting established vision-language capabilities.

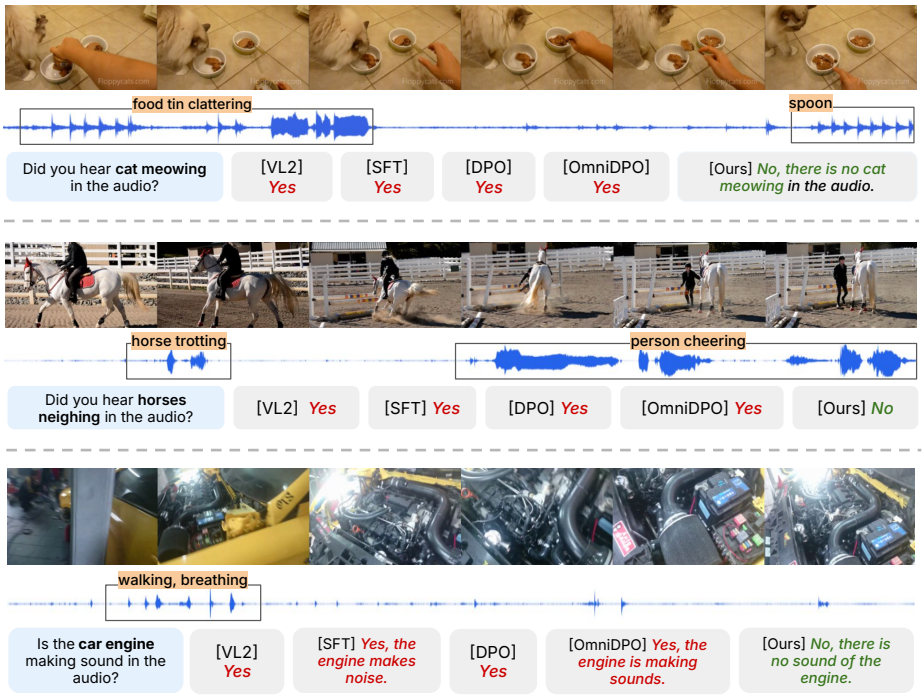

Fig. 4: Qualitative examples of video-driven audio hallucination. Each row shows a video clip (top), its corresponding audio waveform with labeled sound events (middle), and model responses to an audio-focused yes/no question (bottom). In all three cases, the audio contains no evidence of the queried sound, yet all baselines hallucinate affirmative responses. ACPO (Ours) correctly grounds its response in the audio signal.

4 Experiments

4.1 Setup

We describe the benchmarks and datasets we use to validate our approach, overview implementation details and baselines, and present our results.

Benchmarks. AVHBench [43] specifically probes cross-modal hallucination in AVLMs. The benchmark consists of yes/no questions whose ground truth depends on a single modality, while the other may be misleading. We report results on the video-driven audio (V→A) hallucination task, the audio-driven video hallucination (A→V) task, and audio-visual captioning. CMM [23] evaluates unimodal dominance and multimodal robustness under controlled modality masking and corruption. We evaluate on four tasks: Audio-Language, Overreliance on Vision, Overreliance on Audio, and Vision-Audio-Language. We do not include vision-language tasks, since our approach only modifies the audio projector.

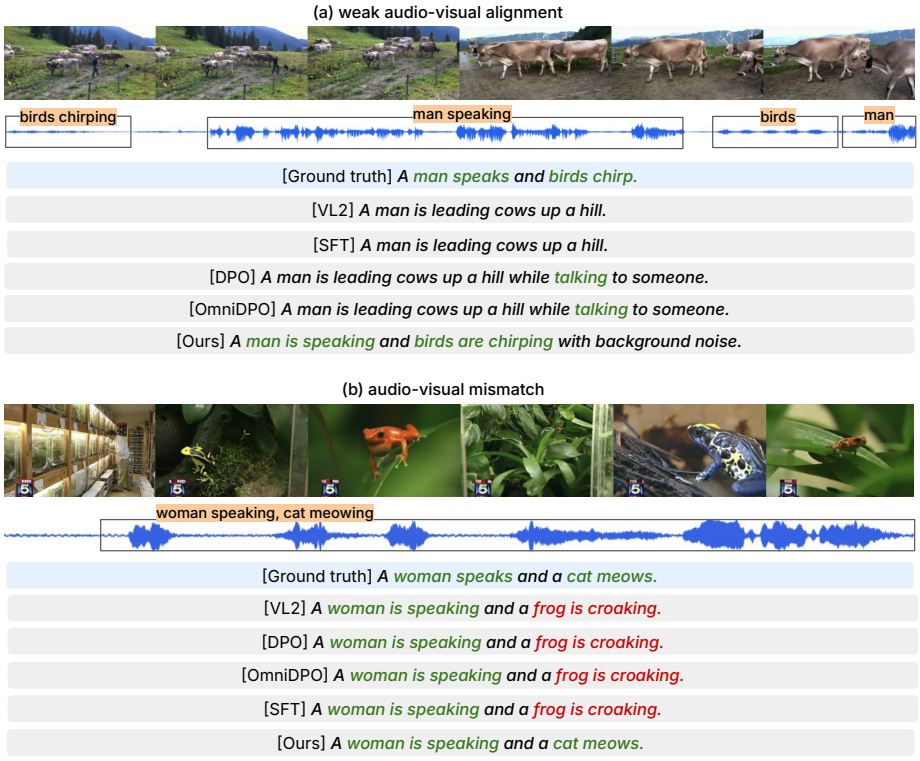

Fig. 5: Qualitative examples of audio-focused captioning. Each row shows a video clip with its labeled audio waveform, a reference audio caption, and model-generated audio captions. (a) The video depicts cows on a hillside, and the audio contains a man speaking and birds chirping. All baselines produce captions grounded in the visual scene ("leading cows up a hill"), failing to describe the actual audio content. ACPO correctly identifies both auditory events. (b) The video shows frogs, but the audio contains a woman speaking and a cat meowing. All baselines hallucinate a frog croaking. ACPO alone correctly describes what is heard.

Unimodal Captioning Evaluation. Open-ended video captioning is highly susceptible to cross-modal hallucination, yet evaluating modality-specific grounding remains a challenge. Standard audio-visual datasets typically provide joint captions that are largely vision-focused. Furthermore, datasets that do offer unimodal annotations [9] often contain generic and uninformative audio descriptions (e.g., "someone is speaking"). This makes it difficult to isolate a model's true audio understanding from its reliance on visual shortcuts. To help close this gap, we constructed a targeted evaluation set to explicitly measure decoupled audio-visual understanding. We sourced videos and their corresponding joint captions from the AVHBench audio-visual captioning task, filtering for 400 clips that contain diverse and distinct audio events. To evaluate robustness against

hallucination, we also constructed 400 audio-swapped counterparts. We used an LLM to rank candidate audio tracks to select plausible mismatches. To generate ground truth audio captions, we provided an AVLM [11] with the original multimodal caption alongside the raw audio track, instructing it to describe only the verifiable auditory events. Conversely, to generate the vision-only ground truth, we prompted the model with the original caption and the raw video frames. We manually verified 10% of the generated captions to confirm factual accuracy and modality attribution. Note that during evaluation, all models receive both video and audio as input. This mirrors the cross-modal hallucination setting: the model must describe what it hears despite the presence of potentially misleading visual input. The unimodal ground-truth captions serve as modality-specific reference targets for measuring grounding. We report METEOR [4] and CIDEr [48] scores for these splits in Table 2.

Training Data. We construct training pairs from 5,000 VALOR [31] clips with joint captions. We decompose each caption into modality-specific targets using GPT-5 [33] and construct audio-swapped pairs as described in Sec. 3.2. We combine our audio-contrastive pairs with the noise-based and text DPO pairs from [7], sampling randomly within each batch at a 60/40 ratio of audio-contrastive to other multimodal pairs.

Implementation Details. We adopt Video-LLaMA2-7B-AV [10] as our pretrained backbone, given its leading performance on the AVHBench and CMM leaderboards. During training, we freeze the video encoder, audio encoder, and the LLM backbone. We fine-tune only the audio projection layer to strengthen audio-language alignment without disrupting the model's established vision-language capabilities. We train for 1 epoch over 5,000 preference pairs using AdamW with a learning rate of 2e-5 and cosine scheduling with warmup, a DPO β of 0.1, and a global batch size of 8. Training completes in approximately 3 hours on a single NVIDIA GH200 120GB GPU.

Baselines. We compare ACPO against the following established training strategies. To ensure a fair comparison, all baseline implementations follow the same training paradigm as ours. Additionally, to provide broader context for our results, we include zero-shot evaluations of several state-of-the-art AVLMs.

1. Base AVLM (zero-shot): Unmodified pretrained Video-LLaMA2-7B-AV [10].
2. SFT [35]: Fine-tuned on VALOR captions using standard cross-entropy loss.
3. DPO [37]: Preference pairs where audio-visual responses are preferred over vision-only responses on original VALOR clips.
4. OmniDPO [7]: Reimplements text preference pairs and noise-based input, adapted to our training data and scope.

We also report zero-shot results for Gemini-Flash-1.5 [45], Qwen2.5-Omni [52], and MiniCPM-o-2.6 [34] as reference points. These models are not directly comparable to our controlled baselines, but provide broader context for situating our results.

Table 1: Correcting audio hallucination on two public benchmarks. Our method achieves the best overall performance on both benchmarks, demonstrating superior audio hallucination resistance and discrimination. PA: accuracy on yes-instances, HR: accuracy on no-instances. Best in bold. *Results reported in cited work.

	AVHBench						CMM						
	Audio Hallucination				Overall		Aud-Lang			Overrely Vision			Overall
Method	Prec↑	Rec↑	F1↑	Acc↑	F1↑	Acc↑	PA↑	HR↑	Acc↑	PA↑	HR↑	Acc↑	Acc↑
Gemini-Flash-1.5* [45]	57.9	94.7	71.9	63.0	77.8	73.2	88.5	39.5	64.0	79.0	36.5	57.8	76.3
Qwen2.5-Omni* [7]	60.8	98.8	75.3	67.6	76.4	70.9	92.0	78.0	85.0	95.0	56.5	75.8	81.0
MiniCPM-o* [7]	70.4	78.6	74.4	72.8	75.1	73.7	95.0	53.0	74.0	91.0	56.5	73.8	76.0
Base model [10]	68.6	88.5	77.3	74.0	77.6	75.6	85.5	85.5	**85.5**	82.0	64.5	73.3	82.5
SFT [35]	72.8	83.5	77.8	76.2	78.1	77.0	84.0	86.0	85.0	80.5	71.5	76.0	82.5
DPO [37]	70.6	**88.9**	78.7	76.0	78.5	76.7	**87.0**	82.5	84.8	**82.5**	72.0	77.3	82.9
OmniDPO [7]	76.6	81.6	79.0	78.4	78.7	78.6	84.5	81.5	83.0	77.0	77.5	77.3	82.4
Ours	**78.2**	82.8	**80.4**	**79.9**	**79.2**	**78.8**	84.5	**86.5**	**85.5**	80.5	**82.0**	**81.3**	**83.4**

Table 2: Unimodal captioning evaluation on our newly introduced eval set. Our method achieves the largest gains on audio captioning while remaining competitive on video, suggesting that improved modality grounding yields richer and more faithful descriptions. Each cell shows METEOR/CIDEr (×100). Original: matched audio-video pairs; Swap: mismatched but plausible substitutions. Best in bold.

Method	Audio (Original)	Audio (Swap)	Video (Original)	Video (Swap)	Average
Base model [10]	18.8/27.9	15.0/13.1	22.7/28.5	**23.2**/29.7	19.9/24.8
SFT [35]	22.4/31.3	17.3/16.9	19.9/**33.6**	20.9/**40.4**	20.1/30.5
DPO [37]	22.4/27.6	18.0/16.5	**22.9**/23.3	23.2/27.1	21.6/23.6
OmniDPO [7]	23.1/33.0	18.9/18.9	21.9/25.1	21.4/26.1	21.3/25.8
Ours	**27.0**/**43.6**	**23.3**/**30.6**	20.4/27.6	20.2/29.7	**22.7**/**32.9**

Each cell: METEOR / CIDEr (×100, ↑).

4.2 Main Results

Audio hallucination. ACPO achieves the strongest audio hallucination performance across both benchmarks (Table 1). On AVHBench audio hallucination, ACPO obtains the highest F1 (80.4) and accuracy (79.9). The baseline exhibits high recall but low precision, indicating a tendency to over-predict plausible sounds. ACPO shifts toward higher precision while maintaining competitive recall. On CMM, ACPO achieves 82.0 hallucination resistance on Overrely Vis, a substantial improvement over the baseline (64.5) and the next best method, OmniDPO (77.5). This subcategory measures robustness when visual content could mislead audio predictions, precisely the failure mode ACPO targets, as Fig. 4 illustrates. ACPO also achieves the highest overall accuracy on both Overrely Vis (81.3) and Aud-Lang (85.5). Aud-Lang evaluates audio understanding without visual input, and ACPO is the only method that does not degrade performance on this subcategory relative to the baseline.

Table 3: Pair type ablation on the AVHBench video-driven audio hallucination task.

Configuration	Acc↑	Prec↑	Rec↑	F1↑
Ours	**79.9**	78.2	82.8	**80.4**
w/o Attribution	79.1	76.2	**84.6**	80.2
w/o Sensitivity	79.3	**78.3**	81.0	79.6

Captioning. ACPO produces the largest gains on audio captioning (Table 2). Audio CIDEr improves from 27.9 to 43.6 on original clips and from 13.1 to 30.6 on swapped clips, substantially outperforming all baselines. These gains demonstrate that ACPO improves the model's ability to describe audio content, not just its ability to reject hallucinated descriptions. The swapped-clip results are particularly informative: when audio and video conflict, ACPO correctly describes what it hears rather than defaulting to visual content, as Fig. 5 illustrates. Video captioning scores also remain competitive with DPO and OmniDPO, decreasing only slightly relative to the baseline, which is expected as the baseline is heavily visually dominant. Overall, ACPO produces a more balanced model that achieves the best average across all conditions on both METEOR and CIDEr.

Multimodal performance. ACPO improves audio grounding without sacrificing broader multimodal capabilities, achieving the best overall accuracy on both AVHBench (78.8) and CMM (83.4). Since ACPO fine-tunes only the audio projection layer, vision-language performance is unaffected by construction. Table 1 reports audio-focused performance; full results on remaining tasks and remaining limitations are provided in the supplementary material.

4.3 Analysis

Ablation. Table 3 validates the contribution of each pair type, with our full model achieving the best accuracy and F1, indicating the best overall balance between precision and recall. Audio-attribution pairs improve precision, reducing the model's tendency to confirm hallucinated sounds, while audio-sensitivity pairs improve recall, reflecting better attention to the actual audio signal. Together they yield the best overall F1. We additionally ablate similarity levels for audio-swapped pairs; results are provided in the supplementary material.

5 Conclusion

In this work, we address the critical challenge of cross-modal hallucination in AVLMs, a pervasive issue where strong visual priors override auditory evidence and cause models to confidently fabricate sounds. To counteract this visual dominance, we propose ACPO, a tailored preference learning framework, which forces the model to decouple spurious audio-visual shortcuts and learn accurate modality attribution. Extensive experiments demonstrate that our approach greatly

mitigates video-driven audio hallucinations while fully preserving the model's overarching multimodal capabilities. Ultimately, this work provides a crucial step toward achieving balanced and trustworthy audio-visual understanding.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 1
2. Ainary, B.: Audo-sight: Enabling ambient interaction for blind and visually impaired individuals. arXiv preprint arXiv:2505.00153 (2025) 1
3. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022) 1
4. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Goldstein, J., Lavie, A., Lin, C.Y., Voss, C. (eds.) Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72. Association for Computational Linguistics, Ann Arbor, Michigan (Jun 2005), https://aclanthology.org/W05-0909/ 10
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) 1
6. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset (2020), https://arxiv.org/abs/2004.14368 3
7. Chen, J., Zhang, T., Huang, S., Niu, Y., Sun, C., Zhang, R., Zhou, G., Wen, L., Hu, X.: Omnidpo: A preference optimization framework to address omni-modal hallucination (2025), https://arxiv.org/abs/2509.00723 3, 4, 10, 11, 18
8. Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., Wei, F.: Beats: Audio pre-training with acoustic tokenizers (2022), https://arxiv.org/abs/2212.09058 3
9. Chen, S., Li, H., Wang, Q., Zhao, Z., Sun, M., Zhu, X., Liu, J.: Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset (2023), https://arxiv.org/abs/2305.18500 9
10. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., Bing, L.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms (2024), https://arxiv.org/abs/2406.07476 1, 3, 10, 11, 18
11. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025), https://arxiv.org/abs/2507.06261 10
12. Dang, J., Deng, S., Chang, H., Wang, T., Wang, B., Wang, S., Zhu, N., Niu, G., Zhao, J., Liu, J.: Hallucination reduction in video-language models via hierarchical multimodal consistency. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence. IJCAI '25 (2025). https://doi.org/10.24963/ijcai.2025/1019, https://doi.org/10.24963/ijcai.2025/1019 4

13. Ge, H., Wang, Y., Yang, M.H., Cai, Y.: Mrfd: Multi-region fusion decoding with self-consistency for mitigating hallucinations in lvlms. arXiv preprint arXiv:2508.10264 (2025) 4
14. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all (2023), https://arxiv.org/abs/2305.05665 7
15. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al.: 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. pp. 14375–14385 (2024) 2, 4
16. Hu, R., Tu, Y., Wei, S., Lu, D., Sang, J.: Prescribing the right remedy: Mitigating hallucinations in large vision-language models via targeted instruction tuning (2025), https://arxiv.org/abs/2404.10332 4
17. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems **43**(2), 1–55 (2025) 2, 4
18. Huh, M., Xue, Z., Das, U., Ashutosh, K., Grauman, K., Pavel, A.: Vid2coach: Transforming how-to videos into task assistants. In: Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology. pp. 1–24 (2025) 1
19. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM computing surveys **55**(12), 1–38 (2023) 2, 4
20. Jung, C., Jang, Y., Chung, J.S.: Avcd: Mitigating hallucinations in audio-visual large language models through contrastive decoding (2025), https://arxiv.org/abs/2505.20862 2, 3, 4, 5
21. Khashabi, D., Min, S., Khot, T., Sabharwal, A., Tafjord, O., Clark, P., Hajishirzi, H.: Unifiedqa: Crossing format boundaries with a single qa system. In: Findings of the association for computational linguistics: EMNLP 2020. pp. 1896–1907 (2020) 1
22. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Advances in neural information processing systems **35**, 22199–22213 (2022) 1
23. Leng, S., Xing, Y., Cheng, Z., Zhou, Y., Zhang, H., Li, X., Zhao, D., Lu, S., Miao, C., Bing, L.: The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio (2024), https://arxiv.org/abs/2410.12787 2, 3, 4, 5, 8
24. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13872–13882 (June 2024) 4
25. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th annual meeting of the association for computational linguistics. pp. 7871–7880 (2020) 1
26. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems **33**, 9459–9474 (2020) 1

27. Li, G., Wei, Y., Tian, Y., Xu, C., Wen, J.R., Hu, D.: Learning to answer questions in dynamic audio-visual scenarios (2022), https://arxiv.org/abs/2203.14072 3
28. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023) 1
29. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 conference on empirical methods in natural language processing. pp. 292–305 (2023) 2, 4
30. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023) 1
31. Liu, J., Chen, S., He, X., Guo, L., Zhu, X., Wang, W., Tang, J.: Valor: Vision-audio-language omni-perception pretraining model and dataset (2025), https://arxiv.org/abs/2304.08345 3, 10
32. Nishimura, T., Nakada, S., Kondo, M.: On the audio hallucinations in large audio-video language models (2024), https://arxiv.org/abs/2401.09774 4
33. OpenAI: Introducing gpt-5. https://openai.com/index/introducing-gpt-5/ (2025) 10
34. OpenBMB: Minicpm-o 2.6. Hugging Face model card: openbmb/MiniCPM-o-2_6 (2025), https://huggingface.co/openbmb/MiniCPM-o-2_6, accessed 2026-03-05 10
35. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022) 10, 11
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021), https://arxiv.org/abs/2103.00020 3
37. Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., Finn, C.: Direct preference optimization: Your language model is secretly a reward model (2024), https://arxiv.org/abs/2305.18290 4, 6, 10, 11, 18
38. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4035–4045 (2018) 2, 4
39. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms (2017), https://arxiv.org/abs/1707.06347 4
40. Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beißwenger, J., Luo, P., Geiger, A., Li, H.: Drivelm: Driving with graph visual question answering. In: European conference on computer vision. pp. 256–274. Springer (2024) 1
41. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P.F.: Learning to summarize with human feedback. Advances in neural information processing systems **33**, 3008–3021 (2020) 1
42. Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., Cai, D.: PandaGPT: One model to instruction-follow them all. In: Hazarika, D., Tang, X.R., Jin, D. (eds.) Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants! pp. 11–23. Association for Computational Linguistics, Prague, Czech Republic (Sep 2023), https://aclanthology.org/2023.tllm-1.2/ 3
43. Sung-Bin, K., Hyun-Bin, O., Lee, J., Senocak, A., Chung, J.S., Oh, T.H.: Avhbench: A cross-modal hallucination benchmark for audio-visual large language models (2025), https://arxiv.org/abs/2410.18325 3, 4, 5, 6, 8

44. Tang, C., Li, Y., Yang, Y., Zhuang, J., Sun, G., Li, W., Ma, Z., Zhang, C.: video-salmonn 2: Caption-enhanced audio-visual large language models (2025), https://arxiv.org/abs/2506.15220 1, 3
45. Team, G.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024), https://arxiv.org/abs/2403.05530 10, 11
46. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 1
47. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025) 3
48. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation (2015), https://arxiv.org/abs/1411.5726 10
49. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022) 1, 4
50. Wu, T.H., Lee, H., Ge, J., Gonzalez, J.E., Darrell, T., Chan, D.M.: Generate, but verify: Reducing hallucination in vision-language models with retrospective resampling (2025), https://arxiv.org/abs/2504.13169 4
51. Xie, Y., Li, G., Xu, X., Kan, M.Y.: V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization (2024), https://arxiv.org/abs/2411.02712 4
52. Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., Zhang, B., Wang, X., Chu, Y., Lin, J.: Qwen2.5-omni technical report (2025), https://arxiv.org/abs/2503.20215 1, 3, 10
53. Yang, P., Wang, X., Duan, X., Chen, H., Hou, R., Jin, C., Zhu, W.: Avqa: A dataset for audio-visual question answering on videos. In: Proceedings of the 30th ACM International Conference on Multimedia. p. 3480–3491. MM '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3503161.3548291, https://doi.org/10.1145/3503161.3548291 3
54. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding (2023), https://arxiv.org/abs/2306.02858 3
55. Zhao, Z., Wang, B., Ouyang, L., Dong, X., Wang, J., He, C.: Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. arXiv preprint arXiv:2311.16839 (2023) 4

Appendix

A Prompt Templates

We provide the prompts used for caption decomposition and evaluation. Variables are shown in `monospace`.

A.1 Training Data: Caption Decomposition

VALOR Caption Splitting (GPT-5)

You split captions into two disjoint captions for a video.

- video_caption: ONLY facts that are explicitly visual. Keep it direct and factual.
- audio_caption: ONLY facts that are explicitly audible. Remove visual details (like colors).

You will receive a JSON array of items as input. For each item, produce an output JSON object:
`{"id":"<same id>", "video_caption":"...", "audio_caption":"..."}`

A.2 Evaluation: Unimodal Caption Generation

Audio Caption Generation (Gemini 2.5 Pro)

You are given an audio clip extracted from a video, along with a reference description: `{caption}`.

Write a single concise sentence describing the sound events in the audio. Name the sounds (e.g., "a dog barks", "bees buzz", "a man speaks") but do NOT transcribe speech or describe sounds in fine detail. Keep it brief.

Video Caption Generation (Gemini 2.5 Pro)

You are given a muted video clip (audio removed), along with a reference description: `{caption}`.

Write a single concise sentence describing what is visible. Focus on the main subjects and actions. Keep it brief.

B Similarity Score Ablation

Table 4 ablates the effect of audio-visual similarity in the swapped pairs, training each pair type in isolation. For audio-attribution pairs, low-similarity swaps yield

Table 4: Similarity score ablation on AVHBench audio hallucination task. Each row trains with a single pair type in isolation. Low-similarity swaps introduce stronger audio-visual conflict; high-similarity swaps produce subtler mismatches. †Video hallucination task accuracy dropped below baseline.

Pair Type	Acc↑	Prec↑	Rec↑	F1↑
Audio-attribution (low sim swap)	**77.7**	76.0	**81.1**	**78.4**
Audio-attribution (high sim swap)	77.0	74.9	**81.1**	77.9
Audio-attribution (no swap)	76.9	**77.6**	75.7	76.6
Audio-sensitivity (low sim swap)†	79.4	**78.9**	80.1	79.5
Audio-sensitivity (high sim swap)†	**79.8**	**78.9**	**81.3**	**80.1**

the best accuracy and F1. Notably, the no-swap variant achieves the lowest accuracy and F1, underscoring that audio-swapped inputs are essential to the effectiveness of attribution pairs. Without them, the audio and video remain aligned, allowing the model to arrive at the preferred audio caption through visual shortcuts alone and weakening the training signal. For audio-sensitivity pairs, the pattern reverses: high-similarity swaps perform best, likely because subtler mismatches produce harder negatives that better expose audio-invariant behavior. However, both sensitivity-only configurations cause a drop in video hallucination accuracy, suggesting that input-contrastive training in isolation can overcorrect toward audio reliance. The full ACPO objective avoids this trade-off.

C Additional Results

Table 5: Additional results on AVHBench and CMM. ACPO remains competitive with other baselines. AV Cap: audio-visual captioning (M: METEOR, C: CIDEr), VAL: vision-audio-language. PA: accuracy on yes-instances, HR: accuracy on no-instances. Best in bold.

	AVHBench						CMM					
	Video Hallucination				AV Cap		VAL			Overrely Audio		
Method	Prec↑	Rec↑	F1↑	Acc↑	M↑	C↑	PA↑	HR↑	Acc↑	PA↑	HR↑	Acc↑
Base model [10]	75.4	80.6	78.0	77.2	17.1	18.3	83.5	96.0	89.8	85.5	84.5	**85.0**
SFT	76.2	81.0	**78.5**	77.9	17.1	18.1	83.5	97.0	90.3	**86.0**	78.0	82.0
DPO [37]	75.3	**81.6**	78.3	77.5	**17.2**	**18.5**	**84.5**	96.5	**90.5**	**86.0**	80.5	83.3
OmniDPO [7]	**79.3**	77.5	78.4	**78.7**	16.8	16.1	79.5	**98.5**	89.0	81.0	**85.5**	83.3
Ours	76.9	79.1	78.0	77.7	**17.2**	18.4	79.5	97.0	88.3	85.0	82.0	83.5

Table 5 reports results on the remaining AVHBench and CMM tasks. We exclude vision-language tasks from CMM entirely, as performance on these tasks

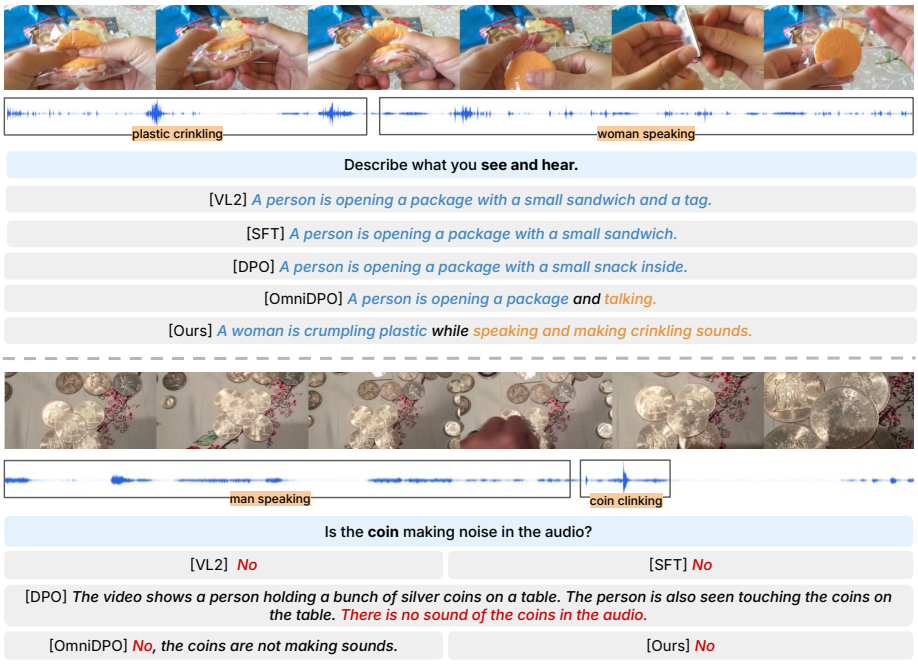

Fig. 6: Qualitative examples illustrating remaining limitations. (Top) Joint audio-visual captioning: ACPO provides the most complete multimodal description but generalizes visual content in favor of auditory detail. (Bottom) Audio-focused question answering: all methods fail to detect a brief, subtle coin clinking sound.

remains identical across all methods by construction, since ACPO only modifies the audio projection layer. On AVHBench, ACPO matches or outperforms the base model on video hallucination and audio-visual captioning, confirming that audio-targeted training does not compromise visual understanding. On CMM overreliance on audio, ACPO exhibits the smallest drop relative to the base model among all preference-based methods. The modest decrease across all methods is expected. The base model's high score on this task partly reflects its tendency to ignore audio altogether, making it trivially robust to misleading audio. As methods improve audio grounding, the model naturally becomes more susceptible to audio interference, as it is actually attending to the audio signal.

D Limitations

Figure 6 illustrates two remaining limitations. In the first example, models are asked to describe what they see and hear. The base model, SFT, and DPO produce purely visual descriptions, ignoring the audio entirely. OmniDPO partially captures audio but drops visual details. ACPO correctly identifies both speech and crinkling sounds. However, in this case, ACPO prioritizes auditory

cues, generalizing visual content in the process (e.g., package with a sandwich $\rightarrow$ crumpling plastic). In the second example, the audio contains a brief coin clinking sound alongside speech. All methods, including ACPO, fail to detect it. This suggests that brief, subtle audio events remain a challenge even with improved audio grounding.