H-MPO: HIERARCHICAL MULTI-RESPONSE PREFERENCE OPTIMIZATION WITH MODALITY-GUIDED DECODING FOR OMNI LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Omni-modal large language models (OLLMs) exhibit capabilities in processing and integrating audio-visual modalities but remain vulnerable to cross-modal hallucinations. To mitigate this, existing methods employ binary preference learning by contrasting relevant responses against completely disconnected ones. This coarse-grained method overlooks intermediate hallucination states, leading to sparse supervision that struggles to address complex cross-modal conflicts. In this paper, we introduce Hierarchical Multi-Response Preference Optimization (H-MPO), which shifts preference learning from a binary comparison to a multi-hierarchy process. To obtain hierarchical preference data, we design an annotation-free Modality-Guided Decoding pipeline that synthesizes a four-level preference ranking set, capturing distinct degrees of modality faithfulness. To leverage the constructed data structure, H-MPO incorporates asymmetric margin constraints to jointly optimize the preference chain. We further propose a token-level reweighting mechanism to strengthen the supervisory signal for modality-sensitive tokens. Extensive experiments on hallucination benchmarks and general tasks demonstrate that H-MPO mitigates cross-modal hallucinations while preserving the general perception and reasoning capabilities of models. Our code and dataset will be released.

1 INTRODUCTION

Recently, Omni-modal Large Language Models (OLLMs) have achieved significant advancements, driving a transition from text-only comprehension to the joint understanding of diverse modalities (Bai et al., 2025; Xu et al., 2025; Ye et al., 2025; Ghosh et al., 2025; Feng et al., 2025). By processing heterogeneous inputs such as video, audio, and text, OLLMs are increasingly capable of perceiving and reasoning about complex environments. However, due to the inherent challenges in multimodal pre-training and alignment, these models remain highly vulnerable to cross-modal hallucinations, manifesting as a critical discrepancy between factual multimodal inputs and generated outputs (Sung-Bin et al., 2024; Leng et al., 2024a; Radevski et al., 2025).

This vulnerability arises from the training of OLLMs, which inevitably absorb dominant unimodal priors and spurious inter-modality correlations from pre-training data. Rather than framing hallucination assessment as a simple binary decision, we systematically categorize model responses into four distinct degrees of modality faithfulness. The first degree represents the ideal, non-hallucinated response, while the subsequent three capture escalating levels of hallucination: (1) *Detailed Description (Non-hallucination)* that faithfully captures fine-grained cues from the question-relevant modality. (2) *Coarse-grained Description* that is correct but lacks modality-specific details. (3) *Modality-confused Interference Description* that is misled by irrelevant modality, blending partial truths. (4) *Complete Modality Disconnect Description* that is driven entirely by irrelevant modality.

Existing methods, *e.g.*, MoD-DPO (Chaubey et al., 2026), mitigate hallucinations by contrasting a fully relevant response (Level 1) with a completely hallucinated one (Level 4). This binary reward design overlooks intermediate states (Levels 2 and 3), leading to sparse supervision signals.

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

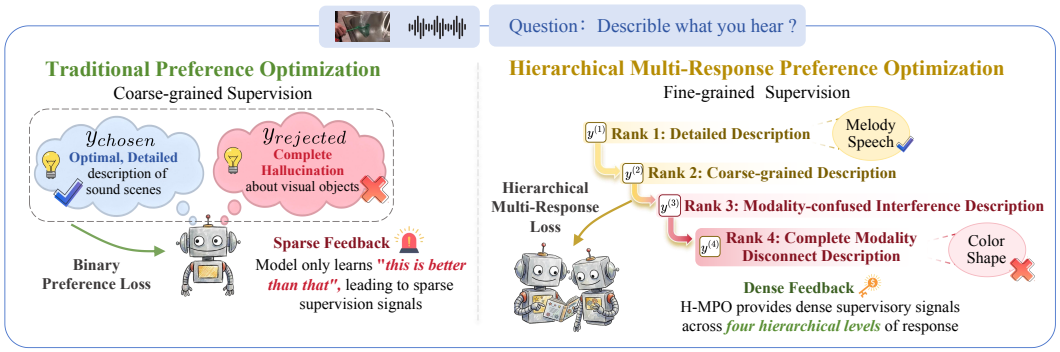


Figure 1: Comparison of traditional preference optimization (left) and our proposed H-MPO (right) under an audio-related question. While traditional methods offer only binary feedback, H-MPO introduces a four-rank preference hierarchy, enabling dense supervision from precise audio descriptions to complete modality disconnection, amplifying factual audio cues (*e.g.*, "melody", "speech") while penalizing hallucinated cross-modal distractors (*e.g.*, "color", "shape").

Consequently, the model struggles to resolve cross-modal conflicts or recover missing details. However, manually constructing such fine-grained preference data in omni-modal scenarios is difficult, as evaluating whether a model missed a visual detail or is biased by irrelevant background audio requires extensive effort and cross-checking.

In this paper, we address these limitations by introducing Hierarchical Multi-Response Preference Optimization (H-MPO). Our core insight is to replace coarse-grained binary comparisons with a fine-grained preference hierarchy that provides explicit supervision across four response tiers. To this end, we design an annotation-free Modality-Guided Decoding pipeline that constructs a four-level preference set without external annotations. By varying input modality combinations and tuning guidance coefficients during decoding, this pipeline leverages the model's own generation capabilities to automatically synthesize preference responses across the four levels of modality faithfulness.

Building upon this hierarchical dataset, we introduce two complementary components to fully exploit its fine-grained structure. The first is Multi-Response Preference Optimization, which extends conventional binary preference optimization by incorporating asymmetric margin constraints across all four responses. This formulation enables joint optimization of the entire preference chain, effectively penalizing hallucinations while preserving factual consistency. The second is Fine-grained Token-level Reweighting strategy, which computes per-token log-probability discrepancies between modality-relevant and modality-irrelevant conditions, thereby amplifying the loss signal on modality-sensitive tokens. As illustrated in Fig. 1, our H-MPO shifts preference learning from a binary decision to a hierarchical supervision.

We evaluate H-MPO on both cross-modal hallucination benchmarks (AVHBench (Sung-Bin et al., 2024) and CMM (Leng et al., 2024a)) and non-hallucination benchmarks (DailyOmni (Zhou et al., 2025) and MMAU (Sakshi et al., 2024)). Experiments demonstrate that H-MPO significantly mitigates cross-modal hallucinations while preserving the fundamental perception and reasoning capabilities. The main contributions of our work are summarized as follows:

- We propose Hierarchical Multi-Response Preference Optimization, transforming binary preference learning into a fine-grained, multi-hierarchy framework that captures four distinct granularities of modality faithfulness.
- We design an annotation-free Modality-Guided Decoding pipeline that enables an automated generation of four-level ranked preference datasets.
- We propose a Multi-Response Preference Optimization to jointly optimize the full preference hierarchy, alongside a token-level reweighting mechanism to amplify critical modality-grounded signals.

- Extensive evaluations demonstrate that H-MPO mitigates cross-modal hallucinations while preserving the model's fundamental perception and reasoning capabilities on unimodal and omni-modal tasks.

2 Related Work

2.1 Hallucination in Multimodal Large Language Models

Hallucination remains an unresolved challenge for multimodal large language models. Early diagnostic benchmarks and methods primarily focused on identifying object-level hallucinations in bimodal settings, attributing failures to language biases (Huang et al., 2024) and modality gaps (Liang et al., 2022; Liu et al., 2024a). However, the transition to omni-modal scenarios introduces more complex cross-modal failure patterns. Advanced benchmarks (Sung-Bin et al., 2024; Leng et al., 2024a) have recently highlighted cross-modal hallucinations, where models misattribute or fabricate evidence across asynchronous visual and auditory streams.

Existing mitigation methodologies generally fall into training-based and training-free paradigms. Training-based approaches update model parameters through additional training to alleviate the illusion problem. This includes the creation of new datasets (Liu et al., 2023; Gunasekar et al., 2023; Chen et al., 2025a; Gunjal et al., 2024; Sun et al., 2023), the design of specialized training objectives to penalize modal inequality (Wang et al., 2024; Xie et al., 2024; Xiao et al., 2024; Wu et al., 2024; 2025), or the application of preference alignment frameworks to multimodal architectures (Chen et al., 2025b; Chaubey et al., 2026). Training-free defenses intervene during the inference phase without updating model weights. These training-free strategies predominantly leverage contrastive decoding mechanisms (e.g., contrasting logits from original and degraded inputs (Chen et al., 2024b; Leng et al., 2024b), or utilizing self-generated descriptions (Chuang et al., 2024; Woo et al., 2024; Huo et al., 2025; Li et al., 2025a)), adjust attention distributions to prioritize visual/auditory evidence, or steer activation spaces during the forward pass (Chen et al., 2024a; Liu et al., 2024b; Jiang et al., 2025; Li et al., 2025b).

2.2 Preference Optimization and Alignment

Direct Preference Optimization (DPO) (Rafailov et al., 2023) and its variants such as Simple Preference Optimization (SimPO) (Meng et al., 2024) offer a rollout-free alternative to reinforcement learning from human feedback (RLHF), aligning language models with factual grounding. However, traditional DPO formulations are fundamentally constrained to the textual domain and cannot be directly transferred to omni-modal scenarios. This limitation arises because text-only DPO optimizes preferences solely based on textual token probabilities, neglecting the complex alignment and cross-modal coordination required among inputs.

To address this limitation, recent works have tailored preference optimization to multimodal and omni-modal settings. Earlier attempts, such as V-DPO (Xie et al., 2024), inject video guidance via a divergence penalty to discourage text-only shortcuts, while mDPO (Wang et al., 2024) enforces modality conditioning by comparing model responses under informative versus degraded visual inputs. To scale these objectives to tri-modal contexts, OmniDPO (Chen et al., 2025b) introduces modality-specific conditional preference learning by constructing text-preference pairs to capture audio-video interactions and multimodal-preference pairs to compel attention toward non-textual streams. More recently, MoD-DPO (Chaubey et al., 2026) proposes a modality-decoupled framework that enforces invariance to irrelevant modalities and sensitivity to relevant ones via KL-based regularizers, coupled with a language-prior debiasing penalty to penalize latent text-only shortcuts.

3 Methodology

3.1 Overview

Given an omni-modal input $X = (x_{\mathrm{V}}, x_{\mathrm{A}}, x_{\mathrm{T}})$, where x_{V} represents the visual stream, x_{A} denotes the auditory signals, and x_{T} is a textual instruction, the objective of OLLMs is to generate a response y that faithfully interprets the multimodal evidence. At each decoding step, the model predicts a

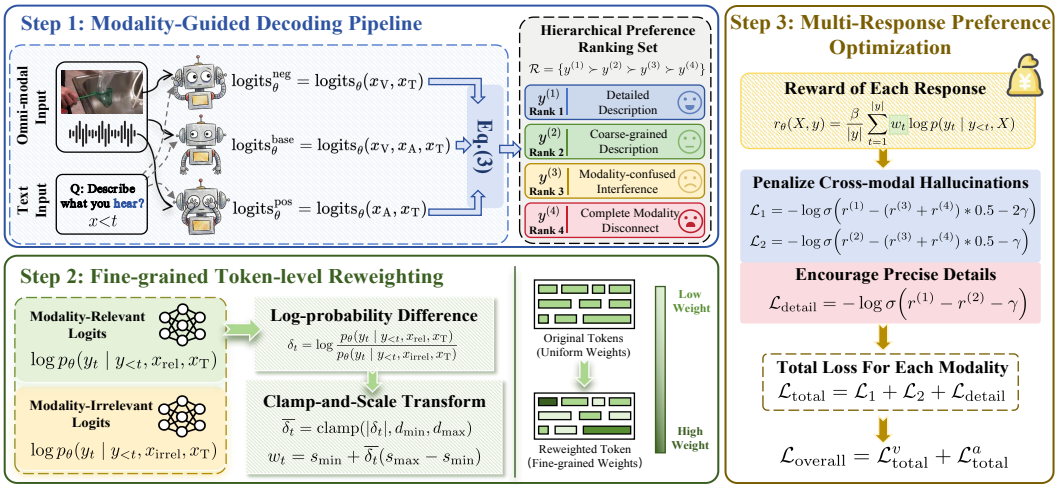


Figure 2: **Overview of the H-MPO framework for mitigating cross-modal hallucinations.** H-MPO consists of three components: **(1) Modality-Guided Decoding Pipeline:** Varies input modality combinations and guidance coefficients during decoding to synthesize an annotation-free four-level preference ranking set R. **(2) Fine-grained Token-level Reweighting:** Computes per-token log-probability differences δ_t between relevant and irrelevant modalities, deriving dynamic weights w_t to emphasize modality-sensitive tokens. **(3) Multi-Response Preference Optimization:** Optimizes the full preference data via asymmetric margin with weighted rewards, effectively penalizing hallucinations while reinforcing fine-grained details.

probability distribution,

$$p(y|x_\text{V}, x_\text{A}, x_\text{T}) = \text{softmax}\left(\text{logits}_\theta(x_\text{V}, x_\text{A}, x_\text{T})\right), \tag{1}$$

where θ denotes the parameters of the OLLM, and $\text{logits}_\theta(\cdot)$ computes the unnormalized prediction scores over the vocabulary. However, existing OLLMs remain vulnerable to cross-modal hallucinations, where this output distribution is misled by dominant yet irrelevant modality priors.

Current preference optimization methods, such as DPO and SimPO, mitigate hallucinations through binary comparison: preferring a modality-relevant response over a completely irrelevant counterpart. However, this framing oversimplifies the nature of hallucinations, providing sparse learning signals. To capture this nuance, we categorize hallucinations into four levels, ordered by their modality faithfulness: (1) detailed description, (2) coarse-grained description, (3) modality-confused interference description, and (4) complete modality disconnect description, where levels (2) and (3), provide critical intermediate supervision absent from traditional binary comparisons. However, manually constructing such multi-level preference data is extremely expensive. We therefore introduce a **Modality-Guided Decoding** pipeline to automatically construct the four-level ranking set $\mathcal{R} = \{y^{(1)} \succ y^{(2)} \succ y^{(3)} \succ y^{(4)}\}$.

Based on the constructed multi-level data structure, we propose a **Multi-Response Preference Optimization** that extends SimPO with asymmetric margin constraints, jointly optimizing the entire preference sequence. To further steer the training signal toward modality-sensitive tokens, we design a **Fine-grained Token-level Reweighting** that computes per-token log-probability differences to derive dynamic weights w_t. The overall framework is shown in Fig 2.

3.2 Modality-Guided Decoding Pipeline For Dataset Construction

Contrastive Decoding (CD) (Li et al., 2023) is designed to adjust the output distribution by shifting the base distribution $\text{logits}_\theta^{\text{base}}$ away from a reference distribution $\text{logits}_\theta^{\text{ref}}$. Formally, CD can be formulated as a single-directional intervention:

$$\text{logits}_\theta^{\text{total}} = \text{logits}_\theta^{\text{base}} + \alpha(\text{logits}_\theta^{\text{base}} - \text{logits}_\theta^{\text{ref}}) \tag{2}$$

where α controls the penalty applied to the reference.

This standard formulation is constrained to a single contrastive direction. In omni-modal scenarios, cross-modal hallucinations stem from complex interactions among multiple modalities, making a single reference insufficient. To obtain responses with controllable modality faithfulness without relying on costly human annotations, we extend Eq.equation 2 to Modality-Guided Decoding (MGD). Specifically, for an instruction x_{T} targeting a specific modality, *e.g.*, "Describe what you see?", we define a relevant modality input x_{V} and an irrelevant modality input x_{A}. Then we can obtain three decoding references from the same OLLM: $\mathrm{logits}_\theta^{\mathrm{pos}} = \mathrm{logits}_\theta(x_{\mathrm{V}}, x_{\mathrm{T}})$ conditioned only on the relevant modality, $\mathrm{logits}_\theta^{\mathrm{neg}} = \mathrm{logits}_\theta(x_{\mathrm{A}}, x_{\mathrm{T}})$ conditioned only on the irrelevant modality, and $\mathrm{logits}_\theta^{\mathrm{base}} = \mathrm{logits}_\theta(x_{\mathrm{V}}, x_{\mathrm{A}}, x_{\mathrm{T}})$ conditioned on both. By decoupling the contrastive directions, the final logits distribution is formulated as:

$$\begin{aligned}\mathrm{logits}_\theta^{\mathrm{total}} = \mathrm{logits}_\theta^{\mathrm{base}} &+ \alpha_{\mathrm{pos}}(\mathrm{logits}_\theta^{\mathrm{pos}} - \mathrm{logits}_\theta^{\mathrm{base}}) \\ &+ \alpha_{\mathrm{neg}}(\mathrm{logits}_\theta^{\mathrm{base}} - \mathrm{logits}_\theta^{\mathrm{neg}}),\end{aligned} \tag{3}$$

Here, $(\mathrm{logits}_\theta^{\mathrm{pos}} - \mathrm{logits}_\theta^{\mathrm{base}})$ enhances fine-grained cues from the relevant modality. Conversely, $(\mathrm{logits}_\theta^{\mathrm{base}} - \mathrm{logits}_\theta^{\mathrm{neg}})$ suppresses spurious interference from the distractor. The coefficients α_{pos} and α_{neg} control the intensity of this enhancement and suppression, respectively.

Multi-response Generation. By tuning α_{pos} and α_{neg} with different configurations, we can generate multi-response across different levels of modality faithfulness,

- **Rank 1 (Detailed Description):** Set with $\alpha_{\mathrm{pos}} = 1, \alpha_{\mathrm{neg}} = 0$, which yields $\mathrm{logits}_\theta^{\mathrm{total}} = \mathrm{logits}_\theta^{\mathrm{pos}}$. The decoding is fully driven by the question-relevant modality, producing highly detailed and faithful descriptions as the upper-bound reference.
- **Rank 2 (Coarse-grained Description):** Set with $\alpha_{\mathrm{pos}} < 0, \alpha_{\mathrm{neg}} = 0$, which yields $\mathrm{logits}_\theta^{\mathrm{total}} = \mathrm{logits}_\theta^{\mathrm{base}} + \alpha_{\mathrm{pos}}(\mathrm{logits}_\theta^{\mathrm{pos}} - \mathrm{logits}_\theta^{\mathrm{base}})$. The negative coefficient suppresses fine-grained cues from the relevant modality, compelling the base model to fall back on coarse-grained descriptions that remain correct but lack specific details.
- **Rank 3 (Modality-confused Interference Description):** Set with $\alpha_{\mathrm{pos}} < 0, \alpha_{\mathrm{neg}} < 0$. The negative α_{pos} suppresses fine-grained cues from the relevant modality, while the negative α_{neg} amplifies interference from the irrelevant modality, leading to a response that mixes coarse-grained correct information with modality-irrelevant content.
- **Rank 4 (Complete Modality Disconnect Description):** Set with $\alpha_{\mathrm{pos}} = 0, \alpha_{\mathrm{neg}} = -1$, which yields $\mathrm{logits}_\theta^{\mathrm{total}} = \mathrm{logits}_\theta^{\mathrm{neg}}$. The decoding is driven by the question-irrelevant modality, steering the response away from the task-relevant modality.

Dataset Construction. We construct the multi-response preference dataset from the MSRVTT(Xu et al., 2016). After filtering out videos without audio, we randomly select 1,212 clips from the remaining set. Then, for each clip, we construct prompts by pairing it with two modality-specific questions: one targeting the visual content ("Describe what you see?") and the other targeting the auditory content ("Describe what you hear?"). To generate hierarchical preference data, we use the MGD previously introduced to generate four responses that span from detailed modality grounding to complete hallucination. In total, we obtain $1{,}212 \times 2 \times 4 = 9{,}696$ responses, which serve as the training data for the subsequent preference alignment stage. Appendix Figs. 10–15 show concrete outputs for video- and audio-driven queries, where the description quality degrades monotonically from fine-grained grounding (Rank 1) to complete modality disconnection (Rank 4).

3.3 Multi-Response Preference Optimization

We adopt SimPO as base preference optimization formulation, which utilizes the length-normalized log-probability of the response as the implicit reward without requiring a separate reward model. Formally, given the omni-modal input $X = (x_{\mathrm{V}}, x_{\mathrm{A}}, x_{\mathrm{T}})$, the reward for a response y is defined as:

$$r_\theta(X, y) = \frac{\beta}{|y|} \sum_{t=1}^{|y|} \log p(y_t \mid y_{<t}, X) \tag{4}$$

where $|y|$ denotes the response length and β is a temperature hyperparameter that scales the reward margin. The SimPO objective maximizes the margin between the rewards of the chosen (y_c) and rejected (y_r) responses:

$$\mathcal{L}_{\text{SimPO}} = -\log \sigma\Big(r_\theta(X, y_c) - r_\theta(X, y_r) - \gamma\Big) \tag{5}$$

where γ is a margin hyperparameter, and σ is the sigmoid function. This formulation encourages the model to assign higher average log-probabilities to chosen responses than to rejected ones by at least γ.

To fully exploit our constructed four-level preference set, we extend this binary formulation to a multi-response objective. We first formulate two losses that penalize cross-modal hallucinations at different levels:

$$\begin{aligned} \mathcal{L}_1 &= -\log \sigma\Big(r^{(1)} - (r^{(3)} + r^{(4)}) * 0.5 - 2\gamma\Big) \\ \mathcal{L}_2 &= -\log \sigma\Big(r^{(2)} - (r^{(3)} + r^{(4)}) * 0.5 - \gamma\Big), \end{aligned} \tag{6}$$

where $r^{(k)} = r_\theta(X, y^{(k)})$. $\mathcal{L}_1$ uses a larger margin 2γ to enforce a robust boundary between the most faithful modality description $y^{(1)}$ and severe hallucinations, while $\mathcal{L}_2$ serves as an intermediate anchor with a relaxed margin γ, ensuring that even coarse-grained correct descriptions $y^{(2)}$ are prioritized over hallucinated ones. By evaluating $y^{(3)}$ and $y^{(4)}$ via their cluster mean, this formulation enforces robust cluster-level separation between faithful and hallucinated responses while avoiding gradient instability from rigid pairwise constraints on severe hallucinations.

To further encourage the model to distinguish between precise details, we introduce

$$\mathcal{L}_{\text{detail}} = -\log \sigma\Big(r^{(1)} - r^{(2)} - \gamma\Big). \tag{7}$$

This loss contrasts $y^{(1)}$ against $y^{(2)}$, forcing the model to capture the distinction between fine-grained and coarse-grained modality descriptions.

The final objective for each modality is the summation:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_{\text{detail}}. \tag{8}$$

Since each training sample is paired with both a visual and an auditory question, the overall training loss is the sum of the modality-specific losses:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}^v_{\text{total}} + \mathcal{L}^a_{\text{total}}, \tag{9}$$

where $\mathcal{L}^v_{\text{total}}$ and $\mathcal{L}^a_{\text{total}}$ denote the losses computed on visual and auditory prompts, respectively.

3.4 Fine-grained Token-level Reweighting

Standard preference losses assign equal weight to all response tokens. However, tokens differ dramatically in their modality-sensitivity: function words and generic phrases are modality-agnostic, while content tokens, *e.g.*, object names, event descriptors, auditory attributes, are direct markers of modality faithfulness. We propose to amplify the loss signal on modality-sensitive tokens.

Specifically, let $x_{\text{rel}} \in \{x_{\text{V}}, x_{\text{A}}\}$ denote the question-relevant modality and x_{irrel} denote the distractor modality. For each token in response y, we compute the token-level log-probability difference:

$$\delta_t = \log \frac{p_\theta(y_t \mid y_{<t}, x_{\text{rel}}, x_{\text{T}})}{p_\theta(y_t \mid y_{<t}, x_{\text{irrel}}, x_{\text{T}})}. \tag{10}$$

While a positive δ_t indicates that token y_t is supported by x_{rel}, a negative value reflects strong influence from x_{irrel}. Therefore, we take the absolute value $|\delta_t|$ to quantify the modality-sensitivity of token y_t, enabling us to prioritize the optimization of tokens tied to multimodal information.

Directly utilizing the raw importance scores $|\delta_t|$ for loss weighting may introduce significant numerical instability during training. To stabilize training and prevent high-variance outliers in raw $|\delta_t|$ from dominating gradient updates, we map importance scores to a bounded range $w_t \in [s_{\min}, s_{\max}]$ via linear scaling and clipping:

$$w_t = s_{\min} + \text{norm}\Big(\text{clamp}(|\delta_t|, d_{\min}, d_{\max})\Big)\Big(s_{\max} - s_{\min}\Big), \tag{11}$$

Table 1: AVHBench results for Qwen2.5-Omni on the two subsets—Audio-driven Video Hallucination and Video-driven Audio Hallucination. Metrics include Accuracy (Acc.), Precision (Prec.), Recall (Rec.), and F1-score. The final column "Avg. Acc" represents the average Accuracy across both tasks. Bold indicates the best performance, while underline indicates the second-best performance. VCD (Leng et al., 2024b) and ICT (Chen et al., 2024a) are training-free decoding-time interventions, while OmniDPO, SimPO, and our H-MPO are training-based preference optimization methods.

	Audio-driven Video Hallucination				**Video-driven Audio Hallucination**				**Overall**
Model	Acc. (↑)	Prec. (↑)	Rec. (↑)	F1 (↑)	Acc. (↑)	Prec. (↑)	Rec. (↑)	F1 (↑)	Avg. Acc (↑)
Qwen2.5-Omni-3B									
Base	77.78	91.53	61.13	73.31	59.03	73.78	27.85	40.43	68.41
+ SFT	78.82	**94.28**	61.23	74.25	75.57	74.13	**78.48**	76.24	77.18
+ SimPO (Meng et al., 2024)	79.35	93.40	63.01	75.26	75.35	76.34	73.40	74.84	77.35
+ H-MPO (Ours)	**82.08**	93.93	**68.50**	**79.22**	**76.62**	**76.43**	76.90	**76.67**	**79.35**
Qwen2.5-Omni-7B									
Base	82.83	93.47	70.60	80.44	82.31	78.51	89.00	83.42	82.57
+ VCD (Leng et al., 2024b)	77.40	75.94	80.20	77.52	68.11	61.40	**97.52**	77.38	72.76
+ ICT (Chen et al., 2024a)	76.70	73.12	**84.45**	78.58	68.80	62.10	96.50	77.80	72.75
+ OmniDPO (Chen et al., 2025b)	**84.42**	88.87	78.70	**83.47**	77.51	70.40	94.93	80.85	80.97
+ SFT	79.70	**96.92**	61.24	75.05	**84.44**	**84.67**	84.08	84.37	82.07
+ SimPO (Meng et al., 2024)	82.35	96.44	67.08	79.12	83.13	80.90	86.70	83.70	82.74
+ H-MPO (Ours)	83.58	92.97	72.57	81.51	84.13	81.86	87.66	**84.66**	**83.86**

where clamp and norm denote the clamping and normalization operations, respectively. To avoid additional computational overhead during the training phase, the token weights w_t are pre-computed offline using the frozen model during dataset construction.

With w_t, Eq.equation 4 can be reformulated as,

$$r_\theta(X, y) = \frac{\beta}{|y|} \sum_{t=1}^{|y|} w_t \log p(y_t \mid y_{<t}, X). \tag{12}$$

By replacing $\log p(y_t \mid y_{<t}, X)$ with $w_t \log p(y_t \mid y_{<t}, X)$ in the reward function, our framework optimizes the preference hierarchy while focusing on modality-grounded tokens. We visualize representative token-level sensitivities, covering grounded, modality-confused, and fully disconnected responses, in Appendix A.6.

4 Experiments

4.1 Experiments Setup

Evaluation benchmarks and metrics. To assess the effectiveness of our framework in both mitigating cross-modal hallucinations and preserving general understanding capabilities, we evaluate on four benchmarks. **AVHBench** (Sung-Bin et al., 2024) focuses on two cross-modal hallucination subsets: Audio-driven Video Hallucination and Video-driven Audio Hallucination, where we report Accuracy, Precision, Recall, and F1 Score. **CMM** employs paired "yes/no" queries to measure Perception Accuracy (PA) and Hallucination Resistance (HR) across two cross-modal hallucination types: spurious inter-modality correlations and uni-modality overreliance. Moreover, to evaluate both omni-modal and audio-only understanding, we additionally report results on **DailyOmni** (Zhou et al., 2025), a multi-choice audio-visual QA benchmark spanning six task families, and **MMAU** (Sakshi et al., 2024), a multi-choice audio understanding benchmark covering speech, sound, and music domains, using average accuracy for both.

Baselines and Models. We conduct experiments using Qwen2.5-Omni (Xu et al., 2025) as the base model. As baselines, we include training-based methods: SFT (fine-tuned on the top-ranked responses y_1), SimPO (trained on standard binary comparison $y^{(1)} \succ y^{(4)}$), and OmniDPO (Chen et al., 2025b), as well as training-free methods such as VCD (Leng et al., 2024b) and ICT (Chen et al., 2024a).

Table 2: Evaluation results on CMM benchmark across Spurious Inter-modality Correlation, Uni-modality Overreliance, and Overall performance. PA denotes Perception Accuracy, HR denotes Hallucination Resistance and HM denotes Harmonic Mean.

Model	Spurious Inter-modality Correlation						Uni-modality Overreliance						Overall		HM
	VL		AL		VAL		Visual Dom		Audio Dom		Lang Dom		PA	HR	
	PA	HR	PA	HR	PA	HR	PA	HR	PA	HR	PA	HR			
Qwen2.5-Omni-3B															
Base	86.5	95.5	84.5	86.5	85.5	**98.0**	58.5	78.5	70.0	**96.5**	63.5	85.0	74.8	90.0	82.4
+SFT	86.5	**96.5**	**87.0**	87.5	86.0	**98.0**	**72.0**	69.5	70.0	94.5	64.0	83.0	77.6	88.2	82.9
+ SimPO (Meng et al., 2024)	87.0	95.5	84.0	**88.0**	85.5	**98.0**	59.5	**80.0**	70.0	**96.5**	64.5	**86.0**	75.1	**90.7**	82.9
+ H-MPO (Ours)	**87.5**	95.0	**87.0**	86.0	**88.0**	**98.0**	**72.0**	78.0	**77.5**	95.0	**68.0**	83.0	**80.0**	89.2	**84.6**
Qwen2.5-Omni-7B															
Base	89.0	98.0	**93.0**	83.0	92.0	**97.5**	90.5	67.5	81.5	81.5	71.0	83.5	86.2	85.2	85.7
+ VCD (Leng et al., 2024b)	93.5	89.0	90.0	77.5	93.5	90.0	94.5	56.5	93.0	40.0	87.5	75.0	92.0	71.3	81.7
+ ICT (Chen et al., 2024a)	**94.5**	90.0	89.0	78.0	94.0	91.5	93.5	57.0	93.5	39.5	**88.0**	75.0	92.1	71.8	82.0
+ OmniDPO (Chen et al., 2025b)	94.0	89.5	91.5	83.5	**95.5**	94.5	**95.5**	63.5	**95.0**	41.5	**88.0**	75.5	**93.3**	74.7	84.0
+ SFT	89.0	98.0	91.5	81.5	93.5	96.5	92.0	68.0	83.5	77.5	70.5	83.5	86.7	84.2	85.6
+ SimPO (Meng et al., 2024)	89.0	**98.5**	92.5	80.5	94.0	93.0	92.0	64.0	85.0	83.5	71.0	**84.5**	87.3	84.0	85.7
+ H-MPO (Ours)	89.0	**98.5**	92.5	**85.0**	93.5	97.0	90.0	**81.5**	84.5	**86.5**	71.0	84.0	86.8	**88.8**	**87.8**

Table 3: Comparison of the proposed approach on general benchmarks: DailyOmni (audio-visual) and MMAU (audio).

Model	DailyOmni	MMAU			
	Acc	Sound	Music	Speech	Avg.
Qwen2.5-Omni-3B	45.12	57.96	47.90	34.23	46.70
+ SFT	42.94	24.32	23.35	12.01	19.90
+ SimPO	44.61	59.16	50.60	38.70	49.50
+ H-MPO	**49.15**	**61.87**	**51.20**	**43.54**	**52.20**

Implementation Details. We utilize both Qwen2.5-Omni-3B and Qwen2.5-Omni-7B as our backbone models. Specifically, we apply LoRA (rank=64) to Qwen2.5-Omni-3B for parameter-efficient tuning, and perform full fine-tuning on Qwen2.5-Omni-7B. All models are optimized over 3 epochs using a cosine learning rate decay scheduler with a 100-step warmup. The learning rate is set to 1×10^{-5} for LoRA and 1×10^{-6} for full fine-tuning. Following the hyperparameter setup in SimPO (Meng et al., 2024), we set the temperature $\beta = 1$ and the target reward margin $\gamma = 0.5$. The token reweighting hyperparameters are reported in Appendix A.1, and we analyze the sensitivity to γ and to the weight bounding range $[s_{\min}, s_{\max}]$ in Figs. 3 and 4.

For training dataset construction by MGD, we generate four ranked responses controlled by different coefficients $(\alpha_{\text{pos}}, \alpha_{\text{neg}})$. Specifically, Rank 1 uses $(\alpha_{\text{pos}}, \alpha_{\text{neg}}) = (1, 0)$. Rank 2 is generated with $(-0.7, 0)$. For Rank 3, the coefficients are set to $(-0.7, -0.45)$ for visual prompts and $(-0.7, -0.2)$ for audio prompts because of the model's stronger baseline visual capability. Finally, Rank 4 is produced using $(0, -1)$. Additionally, for token-level reweighting, we set $d_{\min} = 0$, $d_{\max} = 5$, $s_{\min} = 0.5$, and $s_{\max} = 1.5$.

4.2 Main Results

Performance on Hallucination Benchmarks. As shown in Tab. 1 and Tab. 2, H-MPO consistently outperforms all training-free and training-based baselines across both AVHBench and CMM benchmarks at 3B and 7B scales. Appendix Fig. 16 visualizes representative side-by-side generations of the base model and our H-MPO. Specifically, on CMM benchmark, H-MPO achieves the highest HM (84.6% at 3B, 87.8% at 7B) and HR. Notably, on the Visual Dominance task, H-MPO significantly boosts HR by 13.5% over SFT at 7B, demonstrating its superior capability in mitigating unimodal overreliance.

Table 4: Ablation studies on preference data and loss components.

(a) Comparison between binary preference baselines and our multi-level preference data, and the impact of the modality-aware token reweighting mechanism (w_t).

Model	AVHBench		CMM		
	Acc.	F1	PA	HR	HM
Qwen2.5-Omni-3B	68.4	56.9	74.8	90.0	82.4
$y^{(1)}$ *vs* $y^{(3)}$	77.2	74.9	76.7	90.8	83.7
$y^{(1)}$ *vs* $y^{(4)}$	76.5	73.5	73.7	**91.0**	82.3
All Responses (Ours)	**79.4**	**78.0**	**80.0**	89.2	**84.6**
w/o w_t	76.2	72.3	73.5	**92.1**	82.8
w/ w_t (Ours)	**79.4**	**78.0**	**80.0**	89.2	**84.6**

(b) Contributions of individual loss terms ($\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_{\text{detail}}$) and asymmetric margin configuration.

Model	AVHBench		CMM		
	Acc.	F1	PA	HR	HM
$\mathcal{L}_1$	77.2	74.5	74.6	90.5	82.5
$\mathcal{L}_2$	76.9	74.7	75.7	89.8	82.8
$\mathcal{L}_{\text{detail}}$	75.4	71.4	73.0	**91.2**	82.1
Same γ	78.6	77.7	**80.4**	87.2	83.8
All Losses (Ours)	**79.4**	**78.0**	80.0	89.2	**84.6**

Performance on General Benchmarks. Tab. 3 presents the evaluation results on DailyOmni and MMAU. While standard SFT suffers from severe performance collapse due to catastrophic forgetting, preference optimization methods avoid this degradation. H-MPO achieves the best results across all metrics, improving upon the base model by +4.03% on DailyOmni and +5.50% on MMAU. This confirms that H-MPO effectively reduces hallucinations without sacrificing core multimodal understanding abilities.

5 Ablation Studies

Efficacy of Multi-level Preference Data. As shown in Tab. 4a, comparing our model against binary preference baselines ($y^{(1)}$ vs $y^{(3)}$ and $y^{(1)}$ vs $y^{(4)}$) reveals that while binary pairs improve over the Base model, they lack the granularity needed to distinguish subtle hallucination severities. Incorporating all preference levels ("All Responses") yields the best performance across all benchmarks, reaching a top HM of 84.6% on CMM. This confirms that multi-level preference data provides richer supervisory signals than standard binary pairs, enabling fine-grained discrimination between correct responses and varying degrees of hallucinated outputs.

Effectiveness of Individual Loss Components. Ablating individual loss terms ($\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_{\text{detail}}$) and the asymmetric margin configuration ("Same γ") demonstrates that removing any component leads to clear performance degradation accroding to Tab. 4b. Single loss terms fail to balance perception accuracy and hallucination resistance, whereas a constant margin fails to reflect varying severity gaps. Combining all loss components with asymmetric margins achieves the optimal balance between perception capabilities and hallucination control.

Impact of Token-level Reweighting. As shown in Tab. 4a, removing the dynamic token-level reweighting mechanism ("w/o w_t") leads to clear performance degradation across all benchmarks. This confirms the necessity of prioritizing modality-sensitive tokens, *e.g.*, key objects and event descriptors, over uninformative function words. By focusing optimization on these critical cross-modal cues, our reweighting strategy achieves finer-grained multimodal alignment and effectively suppresses subtle hallucinations.

6 Conclusion

In this paper, We introduced Hierarchical Multi-Response Preference Optimization, a fine-grained framework for cross-modal hallucinations in OLLMs. H-MPO establishes a four-level preference hierarchy capturing varying modality faithfulness. We design an annotation-free Modality-Guided Decoding pipeline to construct hierarchical preference data, and formulate a Multi-Response Optimization objective with asymmetric margins and token-level reweighting to amplify modality-sensitive signals. Experiments on hallucination and general benchmarks show that H-MPO mitigates cross-modal hallucinations while preserving perception and reasoning.

AI USE STATEMENT

In this work, we used generative AI tools for drafting sections of the paper and retrieval and discovery (e.g., finding related work). We have not used generative AI tools for generating synthetic datasets or proving mathematical claims, and research ideation or execution are not applicable to this work. Additionally, we used generative AI tools for aiding or polishing writing. We have reviewed all AI-assisted work. For example, the retrieved literature was manually checked for relevance and accuracy, and the AI-drafted or polished sections were thoroughly reviewed, edited, and verified for correctness by all authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Ashutosh Chaubey, Jiacheng Pang, and Mohammad Soleymani. Mod-dpo: Towards mitigating cross-modal hallucinations in omni llms using modality decoupled preference optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Cong Chen, Mingyu Liu, Chenchen Jing, Yizhou Zhou, Fengyun Rao, Hao Chen, Bo Zhang, and Chunhua Shen. PerturboLLaVA: Reducing Multimodal Hallucinations with Perturbative Visual Training, March 2025a.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. Ict: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models, 2024a. URL https://arxiv.org/abs/2411.15268.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Chao Sun, Rongzhou Zhang, Guanyu Zhou, Lijie Wen, and Xuming Hu. Omnidpo: A preference optimization framework to address omni-modal hallucination, 2025b. URL https://arxiv.org/abs/2509.00723.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding, June 2024b. http://arxiv.org/abs/2403.00425.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models, March 2024. http://arxiv.org/abs/2309.03883.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in MLLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=a2JTVVvcEl.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=FjByDpDVIO.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks are all you need, 2023.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18135–18143, 2024.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv preprint*, March 2024. URL http://arxiv.org/abs/2311.17911.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-Introspective Decoding: Alleviating Hallucinations for Large Vision-Language Models, March 2025.

Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations, 2025. URL https://arxiv.org/abs/2410.02762.

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

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 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*, pp. 13872–13882, 2024b.

Jiaming Li, Jiacheng Zhang, Zequn Jie, Lin Ma, and Guanbin Li. Mitigating Hallucination for Large Vision Language Model by Inter-Modality Correlation Calibration Decoding, March 2025a.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. 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. 12286–12312, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.687. URL https://aclanthology.org/2023.acl-long.687/.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The Hidden Life of Tokens: Reducing Hallucination of Large Vision-Language Models via Visual Information Steering, February 2025b.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning, 2022. URL https://arxiv.org/abs/2203.02053.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A Survey on Hallucination in Large Vision-Language Models, May 2024a.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. Reducing hallucinations in vision-language models via latent space steering, 2024b. URL https://arxiv.org/abs/2410.15778.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Gorjan Radevski, Teodora Popordanoska, Matthew B. Blaschko, and Tinne Tuytelaars. DAVE: Diagnostic benchmark for audio visual evaluation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=4ZAX1NT0ms.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.18290.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark, 2024. URL https://arxiv.org/abs/2410.19168.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf, 2023. URL https://arxiv.org/abs/2309.14525.

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

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mDPO: Conditional Preference Optimization for Multimodal Large Language Models, June 2024.

Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. Don't miss the forest for the trees: Attentional vision calibration for large vision language models, 2024. URL https://arxiv.org/abs/2405.17820.

Kai Wu, Boyuan Jiang, Zhengkai Jiang, Qingdong He, Donghao Luo, Shengzhi Wang, Qingwen Liu, and Chengjie Wang. Noiseboost: Alleviating hallucination with noise perturbation for multimodal large language models, May 2024. http://arxiv.org/abs/2405.20081.

Zongyu Wu, Yuwei Niu, Hongcheng Gao, Minhua Lin, Zhiwei Zhang, Zhifang Zhang, Qi Shi, Yilong Wang, Sike Fu, Junjie Xu, et al. Lanp: Rethinking the impact of language priors in large vision-language models. *arXiv preprint arXiv:2502.12359*, 2025.

Xin Xiao, Bohong Wu, Jiacong Wang, Chunyuan Li, Xun Zhou, and Haoyuan Guo. Seeing the image: Prioritizing visual correlation by contrastive alignment, 2024. URL https://arxiv.org/abs/2405.17871.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-DPO: Mitigating Hallucination in Large Vision Language Models via Vision-Guided Direct Preference Optimization, November 2024.

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, 2025. URL https://arxiv.org/abs/2503.20215.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5288–5296, Las Vegas, NV, USA, June 2016. IEEE. doi: 10.1109/CVPR.2016.571.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, Yuming Lou, Dong Yang, Zhijian Liu, Yukang Chen, Ambrish Dantrey, Ehsan Jahangiri, Sreyan Ghosh, Daguang Xu, Ehsan Hosseini-Asl, Danial Mohseni Taheri, Vidya Murali, Sifei Liu, Jason Lu, Oluwatobi Olabiyi, Frank Wang, Rafael Valle, Bryan Catanzaro, Andrew Tao, Song Han, Jan Kautz, Hongxu Yin, and Pavlo Molchanov. Omnivinci: Enhancing architecture and data for omni-modal understanding llms. *arXiv preprint arXiv:2510.15870*, 2025.

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

A APPENDIX

A.1 SENSITIVITY ANALYSIS

Sensitivity of Target Reward Margin. To analyze the effect of the target reward margin γ, we evaluate performance across $\gamma \in \{0, 0.5, 1.0, 1.5, 2.0\}$ (with 0.5 as our default). As shown in Figure 3, on AVHBench, performance improves steadily up to $\gamma = 0.5$ and plateaus thereafter. On CMM, we observe a clear trade-off between PA and HR: while larger margins, *e.g.*, $\gamma = 1.5$, boost PA, they cause an undesirable drop in HR. Therefore, $\gamma = 0.5$ is selected as default, as it offers the optimal balance, delivering consistent gains in PA while maintaining robust anti-hallucination capabilities.

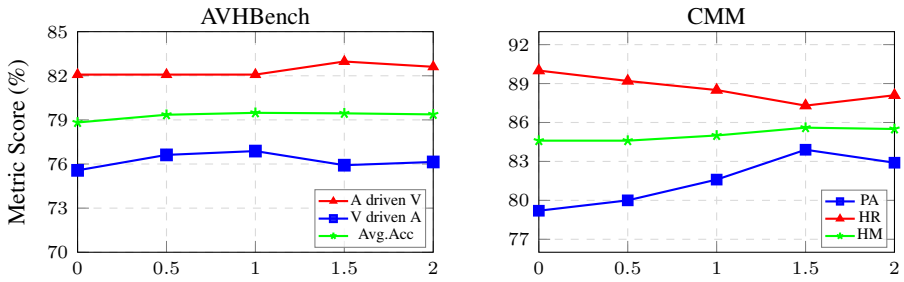


Figure 3: Sensitivity of the target reward margin γ on the AVHBench and CMM benchmark. The y-axis represents the metric score (%).

Sensitivity of Token Weight Bounding Range. To evaluate the impact of the token weight bounding range $[s_{\min}, s_{\max}]$, we compare four configurations: $[0, 1.5]$, $[0.5, 1.5]$ (our default), $[0, 2.0]$, and $[0.5, 2.0]$, as illustrated in Figure 4. Setting a non-zero lower bound ($s_{\min} = 0.5$) is critical: it prevents the over-attenuation of background token gradients, improving HR on CMM. Moreover, expanding the upper bound to $s_{\max} = 2.0$ over-emphasizes specific tokens and degrades PA. Consequently, our default range $[0.5, 1.5]$ achieves the optimal trade-off across both benchmarks.

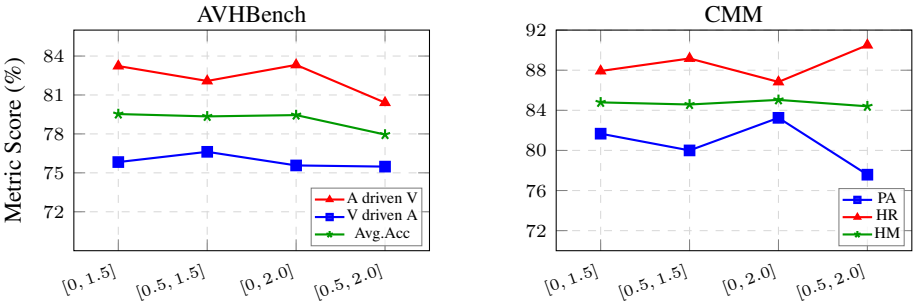


Figure 4: Sensitivity of the token weight bounding range $[s_{\min}, s_{\max}]$ on the AVHBench and CMM benchmarks. The y-axis represents the metric score (%).

A.2 DETAILED TECHNICAL MECHANICS OF MODALITY-GUIDED DECODING

Let $f_\theta(\cdot)$ denote the forward prediction function of the OLLM parameterized by θ. At each auto-regressive decoding step t, given the previously generated sequence $y_{<t}$ and the full omni-modal input $X = (x_{\mathrm{V}}, x_{\mathrm{A}}, x_{\mathrm{T}})$, $f_\theta(\cdot)$ maps the multimodal context into three unnormalized vocabulary logit distributions:

1. **Base Logits ($\mathrm{logits}_\theta^{\mathrm{base}}$):** Computed by conditioning on the full tri-modal context to capture the standard joint prediction, formulated as $\mathrm{logits}_\theta^{\mathrm{base}} = f_\theta(y_{<t}, x_{\mathrm{V}}, x_{\mathrm{A}}, x_{\mathrm{T}})$.

2. **Positive Logits ($\text{logits}_\theta^{\text{pos}}$):** Computed by conditioning strictly on the task-relevant modality $x_{\text{rel}} \in \{x_{\text{V}}, x_{\text{A}}\}$ and prompt x_{T}, isolating uncorrupted target evidence: $\text{logits}_\theta^{\text{pos}} = f_\theta(y_{<t}, x_{\text{rel}}, x_{\text{T}})$.
3. **Negative Logits ($\text{logits}_\theta^{\text{neg}}$):** Computed by conditioning strictly on the distractor modality $x_{\text{irrel}} \in \{x_{\text{V}}, x_{\text{A}}\} \setminus \{x_{\text{rel}}\}$ and prompt x_{T}, capturing spurious cross-modal interference: $\text{logits}_\theta^{\text{neg}} = f_\theta(y_{<t}, x_{\text{irrel}}, x_{\text{T}})$.

By decoupling the positive enhancement and negative suppression directions, the unified token distribution is governed by the fused logit tensor $\text{logits}_\theta^{\text{total}}$:

$$\begin{aligned}\text{logits}_\theta^{\text{total}} = \text{logits}_\theta^{\text{base}} &+ \alpha_{\text{pos}}(\text{logits}_\theta^{\text{pos}} - \text{logits}_\theta^{\text{base}}) \\ &+ \alpha_{\text{neg}}(\text{logits}_\theta^{\text{base}} - \text{logits}_\theta^{\text{neg}}),\end{aligned} \tag{13}$$

where α_{pos} and $\alpha_{\textbf{neg}}$ dynamically control the strength of target feature amplification and distractor noise suppression, respectively. **Notably, since all three distributions are generated by the same underlying model f_θ, they naturally share an identical vocabulary space, inherently avoiding token alignment errors or vocabulary mismatch issues commonly encountered in multi-model contrastive decoding.**

To prevent memory bottlenecks and $3\times$ latency overhead during auto-regressive decoding, our implementation maintains isolated Key-Value (KV) Caches and dynamic attention masks for auxiliary models across GPU devices. During the **Prefill stage** ($t = 1$), full multimodal representations (video frames, audio spectrograms, and text tokens) are processed independently to initialize respective KV-caches. In subsequent **Decoding stages** ($t > 1$), only the newly generated token ID is transferred to auxiliary GPUs, appending directly to the cached key-value states before logits are gathered back to the primary GPU.

A.3 Experiment Details

For dataset construction, all generation procedures were executed across $3 \times$ NVIDIA RTX PRO 6000. During our experiments, we used the following inference parameters to ensure standard model outputs: a temperature of 0.7, a top-p value of 0.9 with nucleus sampling enabled (`do_sample=True`), and a maximum response budget of max_new_tokens $= 128$.

A.4 Qualitative Examples of Four-Level Preference Dataset

To demonstrate the objective validity and consistency of the four-level preference hierarchy ($y^{(1)} \succ y^{(2)} \succ y^{(3)} \succ y^{(4)}$), we provide a fine-grained comparative analysis across video-driven queries (Figures 10–12) and audio-driven queries (Figures 13–15)

Across both video-driven and audio-driven prompts, the MGD pipeline enforces a strictly monotonic degradation in modality faithfulness:

1. **Rank 1 vs. Rank 2 (Granularity Gap):** As shown in the audio-related task (Figure 13), Rank 1 precisely captures fine-grained, highly specific semantic and acoustic details (*"a change of plans to check out with the tires, a reference to a Lamborghini... sister going inside"*). In contrast, Rank 2 remains factually correct but degrades into a coarse, generic summary (*"a man is speaking while a car engine is idling"*), confirming that Rank 1 offers superior information density ($y^{(1)} \succ y^{(2)}$).
2. **Rank 2 vs. Rank 3 (Cross-Modal Contamination):** In the video-related task (Figure 10, "Describe what you see?"), Rank 2 correctly describes the visual environment (*"small animal... mountainous landscape with a river"*). Rank 3 retains part of this visual content but becomes contaminated by irrelevant auditory background cues, falsely hallucinating *"a vibrant music track in the background"* (red text). This injects spurious cross-modal interference while maintaining partial ground-truth, placing it strictly below Rank 2 ($y^{(2)} \succ y^{(3)}$).
3. **Rank 3 vs. Rank 4 (Complete Evidence Disconnect):** In Rank 4, task-relevant modality grounding completely breaks down. As illustrated in Figure 13, Rank 4 fabricates entirely

Model	AVHBench		CMM		
	Acc.	F1	PA	HR	HM
Qwen2.5-Omni-3B	68.4	56.9	74.8	90.0	82.4
Decomposed Loss	74.1	67.8	69.3	**93.3**	81.3
H-MPO (Ours)	**79.4**	**78.0**	**80.0**	89.2	**84.6**

Table 5: Performance comparison between the Decomposed Pairwise Hallucination Baseline ($\mathcal{L}_{\text{decomposed}}$) and our H-MPO objective ($\mathcal{L}_{\text{total}}$) on Qwen2.5-Omni-3B.

false auditory attributes (*"a man speaking in Arabic... ambient sounds of nature"*), completely hallucinating the language and acoustic environment while ignoring the actual English speech and automotive context. This establishes Rank 4 as the absolute lower-bound hallucination state ($y^{(3)} \succ y^{(4)}$).

These qualitative transitions confirm that the logit-level intervention mechanically controls the model's output distribution, yielding a reliable, structurally ordered preference hierarchy across diverse omni-modal queries.

A.5 Effectiveness of Cluster-Averaged Hallucination Penalty

In our proposed H-MPO framework, the training objective consists of two functional parts: a fine-grained loss $\mathcal{L}_{\text{detail}}$ to distinguish precise details ($y^{(1)} \succ y^{(2)}$), and hallucination penalty losses ($\mathcal{L}_1, \mathcal{L}_2$) to enforce boundaries between non-hallucinated and hallucinated states. Crucially, in $\mathcal{L}_1$ and $\mathcal{L}_2$, we represent the hallucination target as the cluster mean of the hallucinated rewards, i.e., $\frac{r^{(3)}+r^{(4)}}{2}$, rather than unrolling them into individual pairwise comparisons.

Formulation of Decomposed Pairwise Hallucination Baseline. To justify the necessity of averaging hallucination rewards over decomposing them into explicit pairwise loss terms, we construct a Decomposed Pairwise Hallucination Baseline. In this baseline, the fine-grained loss $\mathcal{L}_{\text{detail}} = -\log\sigma(r^{(1)} - r^{(2)} - \gamma)$ remains unchanged, but the cluster-averaged hallucination losses $\mathcal{L}_1$ and $\mathcal{L}_2$ are replaced by unrolled pairwise losses $\mathcal{L}'_1$ and $\mathcal{L}'_2$:

$$\begin{aligned}\mathcal{L}'_1 &= -\log\sigma(r^{(1)} - r^{(3)} - 2\gamma) - \log\sigma(r^{(1)} - r^{(4)} - 3\gamma),\\ \mathcal{L}'_2 &= -\log\sigma(r^{(2)} - r^{(3)} - \gamma) - \log\sigma(r^{(2)} - r^{(4)} - 2\gamma).\end{aligned} \tag{14}$$

The total loss for this decomposed baseline is computed as the direct summation of the detail loss and the two unrolled hallucination losses:

$$\mathcal{L}_{\text{decomposed}} = \mathcal{L}_{\text{detail}} + \mathcal{L}'_1 + \mathcal{L}'_2. \tag{15}$$

Experimental Results and Analysis. As shown in Table 5, comparing the Decomposed Loss baseline with our proposed H-MPO framework reveals several critical empirical insights:

1. **Perception Collapse in Decomposed Pairwise Losses:** While the Decomposed Loss achieves a high Hallucination Resistance (HR $= 93.3\%$), it suffers from severe perception degradation. Its Perception Accuracy on CMM drops to 69.3%, which is not only 10.7% lower than H-MPO (80.0%) but even 5.5% below the unaligned base model (74.8%). This indicates that unrolling explicit, increasing pairwise margins ($2\gamma, 3\gamma$) against individual hallucinated targets severely over-penalizes the reward space, forcing the model to become overly conservative and frequently decline to output factual descriptions. As a result, its overall Harmonic Mean on CMM (HM $= 81.3\%$) drops below the base model (82.4%).
2. **Superiority of Cluster-Averaged Macro-Boundaries:** In contrast, H-MPO leverages the cluster mean $\frac{r^{(3)}+r^{(4)}}{2}$ in $\mathcal{L}_1$ and $\mathcal{L}_2$ to treat $y^{(3)}$ and $y^{(4)}$ as a collective hallucination target space. This formulation establishes a smooth macro-boundary between non-hallucinated and hallucinated states without getting trapped in rigid, individual pairwise margin constraints. Consequently, H-MPO achieves substantial improvements across all key perception and alignment metrics (79.4% Acc and 78.0% F1 on AVHBench; 80.0% PA and 84.6%

HM on CMM), demonstrating that cluster-averaged reward aggregation effectively suppresses cross-modal hallucinations while preserving core multimodal perception.

A.6 Fine-grained Token-Level Weight Visualization and Case Analysis

To qualitatively demonstrate our Fine-grained Token-level Reweighting mechanism can accurately identify modality-sensitive tokens, we visualize token-level modality sensitivity (δ_t) representative generation cases in Figures 5 to 9.

In accordance with Eq. (10), the token-level modality sensitivity originates from the raw log-probability difference:

$$\delta_t = \log \frac{p_\theta(y_t \mid y_{<t}, x_{\text{rel}}, x_{\text{T}})}{p_\theta(y_t \mid y_{<t}, x_{\text{irrel}}, x_{\text{T}})}, \tag{16}$$

where x_{rel} denotes the target modality relevant to the instruction x_{T}, and x_{irrel} represents the interfering distractor modality. While the preference training objective scales $|\delta_t|$ into $w_t \in [s_{\min}, s_{\max}]$ via Eq. (11) to stabilize gradient updates, for qualitative case analysis, we directly visualize the raw, unprocessed difference δ_t. Positive values ($\delta_t > 0$) indicate strong target-modality grounding and are highlighted in **Green**, whereas negative values ($\delta_t < 0$) signify that token prediction is misled by the distractor modality or unimodal priors, highlighted in **Red**.

We categorize our visualization analysis into three distinct modality faithfulness scenarios:

1.Target-Modality Driven Descriptions. Figures 5 and 6 showcase responses where generated tokens strictly depend on the target modality evidence ($\delta_t > 0$).

- **Audio Modality Case (Figure 5):** Under an auditory query (*"Describe what you hear?"*), our mechanism assigns high positive values δ_t to specific acoustic keywords such as `"amidst"` ($\delta_t = +2.90$), `"man"` ($\delta_t = +1.58$), `"siren"` ($\delta_t = +1.25$), and `"crowd"` ($\delta_t = +0.98$). Modality-agnostic function words (e.g., `"with"`, `"towards the end"`) yield $\delta_t \approx 0$ and remain uncolored.
- **Visual Modality Case (Figure 6):** Under a video query (*"Describe what you see?"*), content-dense visual markers—such as `"secured"` ($\delta_t = +8.53$), `"young"` ($\delta_t = +8.50$), and `"graphic"` ($\delta_t = +4.49$)—receive near-maximum dark green highlights, confirming their strong reliance on visual inputs rather than audio background noise.

2. Completely Disconnected Descriptions (Irrelevant Modality Overreliance). Figures 7 and 8 present severe Rank 4 hallucination cases where the model outputs details entirely derived from the irrelevant distractor stream.

- **Audio Query with Visual-only Output (Figure 7):** When asked *"Describe what you hear?"*, the model erroneously describes visual actions (`"kiss"`, `"sunlight"`, `"pool"`). Because these tokens receive higher likelihood conditioned on x_{irrel} (video) than on x_{rel} (audio), they yield significantly negative log-probability differences ($\delta_t \ll 0$) and are flagged entirely in **Red**.
- **Video Query with Audio-only Output (Figure 8):** When prompted with *"Describe what you see?"*, the model generates pure acoustic descriptors (`"audio"`, `"engine"`, `"wind"`). Consequently, the entire sequence exhibits negative δ_t values and is shaded in **Red**.

3. Modality-Confused Descriptions. Figure 9 illustrates a subtle Rank 3 hallucination state where the model mixes true visual grounding with auditory distractors under a visual query.

- **Mixed Case Analysis (Figure 9):** Grounded visual concepts (`"natural"`, `"river"`, `"mountains"`) exhibit positive $\delta_t > 0$ (**Green**), whereas spurious auditory hallucinated attributes (`"vibrant"`, `"music"`) yield negative $\delta_t < 0$ (**Red**) due to background audio cues. This mixed visualization proves that our token-level discrepancy metric δ_t can accurately locate token-level hallucination span within an otherwise plausible sentence.

The audio features a man speaking amidst background crowd noise and cheering, with a siren heard towards the end.

Figure 5: Token-level weight visualization under an audio-related prompt (*"Describe what you hear?"*). Darker green highlights represent higher positive raw log-probability differences ($\delta_t > 0$), indicating strong acoustic grounding.

In the video, there are two young men sitting in the front seats of a car. They are both wearing white t-shirts with graphic designs and are secured with seat belts.

Figure 6: Token-level weight visualization under a video-related prompt (*"Describe what you see?"*). Darker green highlights correspond to higher positive raw log-probability differences ($\delta_t > 0$), indicating strong visual grounding.

In the first scene, two individuals kiss passionately, surrounded by sunlight streaming through a window. The setting then transitions to an outdoor pool area where a woman in a yellow outfit bends forward, shielding her face with her hands.

Figure 7: Token-level weight visualization under an audio prompt (*"Describe what you hear?"*) for a completely disconnected response (Rank 4). Because the response contains purely visual descriptions, content tokens yield negative log-probability differences ($\delta_t \ll 0$) and are highlighted in **Red**.

The audio features a man speaking, likely in a car, as indicated by the sound of the engine and wind.

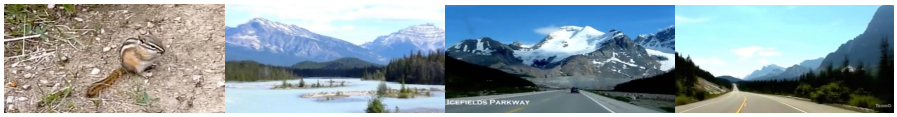
Figure 8: Token-level weight visualization under a video prompt (*"Describe what you see?"*) for a Rank 4 response. The model outputs exclusively auditory details, causing the sentence tokens to yield negative discrepancies ($\delta_t \ll 0$), highlighted in **Red**.

The scene presents a serene and beautiful natural environment, featuring mountains, a river, and a vibrant music track in the background.

Figure 9: Token-level weight visualization under a video prompt (*"Describe what you see?"*) for a modality-confused response (Rank 3). **Green** highlights mark target-relevant visual grounding ($\delta_t > 0$), whereas **Red** highlights expose hallucinated auditory distractors ($\delta_t < 0$).

A.7 Qualitative Results

As shown in Figure 16, we provide qualitative examples showing how our method mitigates cross-modal hallucinations.



Prompt: Describe what you see? Video-related

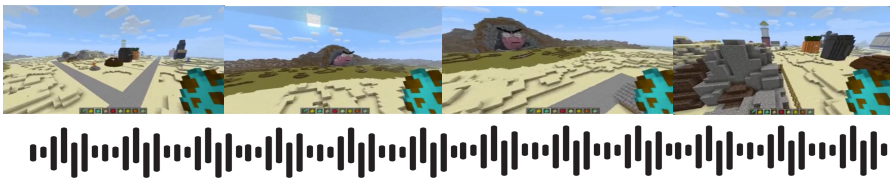
Rank 1 (Detailed Description)
The video begins with a close-up shot of a bird's nest on the ground. The nest appears to be made of twigs and other natural materials. The camera then pans out to reveal a stunning mountain landscape with snow-capped peaks and a winding road. The road is surrounded by dense forests and the sky is clear and blue. The camera follows the road as it winds through the mountains, providing a breath-taking view of the natural beauty of the area. The video ends with a shot of the road disappearing into the distance, leaving the viewer with a sense of awe and wonder at the majesty of nature.

Rank 2 (Coarse-grained Description)
The image first shows a small animal, possibly a squirrel, on the ground, then transitions to a stunning view of a mountainous landscape with a river flowing through it. The scene shifts to a road surrounded by mountains, and finally, it moves to a wide road in the mountains.

Rank 3 (Modality-confused Interference Description)
The scene presents a serene and beautiful natural environment, featuring mountains, a river, and a vibrant music track in the background.

Rank 4 (Complete Modality Disconnect Description)
The image shows a guitar being played, likely on a fretboard, with the neck of the guitar visible and the strings being strummed.

Figure 10: Video Case 1



Prompt: Describe what you see? Video-related

Rank 1 (Detailed Description)
The image depicts a scene from the popular sandbox video game Minecraft. The player is navigating through a desert biome, characterized by sandy terrain and sparse vegetation. The sky is clear with a few clouds, and the player is holding a sword, indicating they are prepared for potential threats. In the distance, there is a structure made of various blocks, possibly a base or a village. The landscape includes a mix of sand, dirt, and scattered rocks, typical of a desert environment.

Rank 2 (Coarse-grained Description)
In the video, I see a Minecraft game where the player navigates through a desert biome, encountering various structures and landscapes. The player's inventory is also visible, showing items like a sword and blocks.

Rank 3 (Modality-confused Interference Description)
The image contains a giant worm from an episode hated during childhood, appearing scary due to its size and appearance.

Rank 4 (Complete Modality Disconnect Description)
The image shows a person wearing a black jacket with a white design on the front, standing against a plain background.

Figure 11: Video Case 2



Prompt: Describe what you see? Video-related

Rank 1 (Detailed Description)
The video shows a rugby match in progress. A player in a white jersey is seen running and then tackling another player. The ball is then passed to a player in a purple jersey, who continues to run with it.

Rank 2 (Coarse-grained Description)
In this video, a man wearing a white shirt is seen running on a field with a ball in his hand. He runs towards another player who is also holding a ball. The two players then start running towards the opposite direction, with one player chasing the other.

Rank 3 (Modality-confused Interference Description)
The audio contains a mix of sounds including music, video game sounds, and cheering, interspersed with male speech.

Rank 4 (Complete Modality Disconnect Description)
In the audio, there is a speech synthesizer speaking over a radio, likely providing updates or commentary on a sports game, with mentions of England and New Zealand and a score of 18 to 14. The scene is accompanied by music, and the synthesizer speaks again at the end.

Figure 12: Video Case 3



Prompt: Describe what you hear? Audio-related

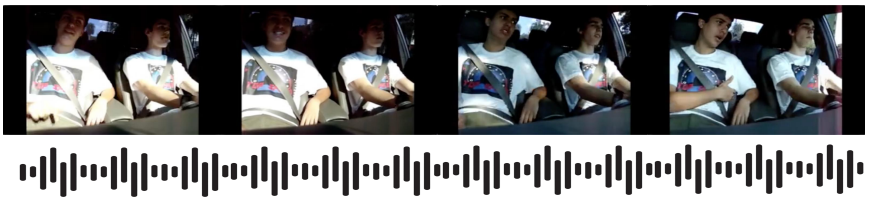
Rank 1 (Detailed Description)
The audio mentions a change of plans to check out with the tires, a reference to a Lamborghini, and a mention of a sister going inside, although the speaker is unsure of what she is doing.

Rank 2 (Coarse-grained Description)
In the audio, a man is speaking while a car engine is idling in the background.

Rank 3 (Modality-confused Interference Description)
The audio mentions that the speaker's plan is to go and get the car, which is likely the white sports car seen in the video. The speaker is wearing a green polo shirt and a cap, and is talking to the camera while walking next to the car. The audio also mentions that the speaker's friend will get a new car.

Rank 4 (Complete Modality Disconnect Description)
The audio features a musical piece that is energetic and lively, with a fast tempo and a catchy melody. The instrumentation includes a variety of instruments, including electric guitar, drums, and keyboards, creating a dynamic and engaging sound. The lyrics are upbeat and positive, with a focus on personal growth and self-improvement.

Figure 13: Audio Case 1



Prompt: Describe what you hear? Audio-related

Rank 1 (Detailed Description)
The audio features a woman speaking in Portuguese, with the content being unclear and difficult to understand.

Rank 2 (Coarse-grained Description)
The audio features a car's interior sounds and a woman singing in the background.

Rank 3 (Modality-confused Interference Description)
Two boys sit in a car, smile, and talk. The driver turns his head from side to side.

Rank 4 (Complete Modality Disconnect Description)
The video shows two young men in a car, with one driving and the other sitting in the passenger seat. They appear to be engaged in a conversation, with the passenger making gestures and the driver occasionally looking at him. The car's interior is visible, including the steering wheel and dashboard. The background outside the car shows a sunny day with trees and buildings passing by.

Figure 14: Audio Case 2



Prompt: Describe what you hear? Audio-related

Rank 1 (Detailed Description)
In this audio, a male voice speaks with a hint of laughter in the background, and the speaker reminisces about a giant worm from an episode that they disliked as a child.

Rank 2 (Coarse-grained Description)
The audio features a male voice speaking in a casual tone, and there is a background sound that appears to be video game music or sound effects.

Rank 3 (Modality-confused Interference Description)
The audio features speech in English, likely discussing Minecraft gameplay and resources. There is also background music playing.

Rank 4 (Complete Modality Disconnect Description)
The audio features the ambient sounds of a Minecraft game, including the player's movements and interactions with the environment. The game's music and sound effects can be heard, adding to the immersive experience.

Figure 15: Audio Case 3

Prompt: Is the crowd making sound in the audio?



※Video-driven Audio Hallucination

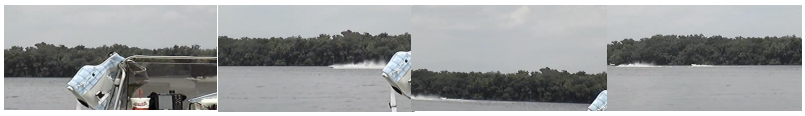
Base Model

Yes, the crowd is making sound in the audio. It's a bit of a noisy environment. What do you think about it?

H-MPO

No, the crowd isn't making sound in the audio it's just music so, what do you think about the music?

Prompt: Is the distant machine visible in the video?



※Audio-driven Video Hallucination

Base Model

Yes, the distant machine is visible in the video. It's making a lot of noise. What do you think it might be?

H-MPO

No, i don't see anything that looks like a machine in the distance. Did you see something that looked like a machine? let's keep chatting about it.

Figure 16: Qualitative Results in Qwen2.5-Omni