# OMNIDPO: A Preference Optimization Framework to Address Omni-Modal Hallucination

**Junzhe Chen**[*1], **Tianshu Zhang**[*1], **Shiyu Huang**[2], **Yuwei Niu**[3], **Chao Sun**[1], **Rongzhou Zhang**[1], **Guanyu Zhou**[4], **Lijie Wen**[†1]

[1] Tsinghua University
[2] OpenRL
[3] Chongqing University
[4] The Hong Kong University of Science and Technology (Guangzhou)
chenjz24@mails.tsinghua.edu.cn
wenlj@tsinghua.edu.cn

## Abstract

Recently, Omni-modal large language models (OLLMs) have sparked a new wave of research, achieving impressive results in tasks such as audio-video understanding and real-time environment perception. However, hallucination issues still persist. Similar to the bimodal setting, the priors from the text modality tend to dominate, leading OLLMs to rely more heavily on textual cues while neglecting visual and audio information. In addition, fully multimodal scenarios introduce new challenges. Most existing models align visual or auditory modalities with text independently during training, while ignoring the intrinsic correlations between video and its corresponding audio. This oversight results in hallucinations when reasoning requires interpreting hidden audio cues embedded in video content. To address these challenges, we propose OMNIDPO, a preference-alignment framework designed to mitigate hallucinations in OLLMs. Specifically, OMNIDPO incorporates two strategies: (1) constructing text-preference sample pairs to enhance the model's understanding of audio-video interactions; and (2) constructing multimodal-preference sample pairs to strengthen the model's attention to visual and auditory information. By tackling both challenges, OMNIDPO effectively improves multimodal grounding and reduces hallucination. Experiments conducted on two OLLMs demonstrate that OmniDPO not only effectively mitigates multimodal hallucinations but also significantly enhances the models' reasoning capabilities across modalities.

## Introduction

Omni-modal large language models (OLLMs) have rapidly advanced in their ability to understand and generate content from combined inputs such as images, audio, and text. By leveraging information from multiple modalities, these models can perform complex tasks ranging from video question answering to audio-visual scene description. However, despite these advancements, hallucination remains a critical issue: models often produce outputs that do not accurately reflect the given visual or auditory input (Nishimura, Nakada, and Kondo 2024; Sahoo et al. 2024). This phenomenon of generating inconsistent or fabricated omni-modal content is broadly referred to as omni-modal hallucination. It poses serious risks in high-stakes applications (e.g. autonomous driving (Wang 2024)), where factual precision is paramount.

Previous studies in the bimodal (text-image) setting have shown that the significantly stronger capability of the text model compared to the visual model results in dominant textual priors (Bai et al. 2025b; Huang et al. 2024b). Consequently, models tend to rely heavily on input text while overlooking visual information—a primary cause of multimodal hallucinations (Favero et al. 2024). Similarly, OLLMs exhibit comparable issues: they are inclined to depend on textual inputs while neglecting other modalities (Gao et al. 2025). Furthermore, omni-modal scenarios introduce new challenges. Existing OLLMs typically align vision or audio with text separately during training, while failing to account for the intrinsic interactions between video and its associated audio (Guo et al. 2025). This limitation leads to hallucinations, especially in tasks that require reasoning based on subtle audio cues embedded within video content. While the issue of dominant textual priors has been extensively studied in bimodal settings, its manifestation in omni-modal scenarios remains under-explored. To bridge this gap, we conduct a targeted analysis using Qwen2.5-Omni-7B on examples involving both audio and video cues (see Figure 1). Our case study reveals that even in omni-modal contexts, the model frequently ignores audio signals—such as human voices or environmental sounds—when generating answers, instead relying on hallucinated textual priors. This suggests that the problem of textual dominance persists beyond the bimodal setting and must be explicitly addressed in omni-modal learning frameworks.

Research on mitigating multimodal hallucinations generally falls into two categories. 1) Training-free methods, which do not modify the model's parameters. Representative examples include contrastive decoding (e.g., VCD (Leng et al. 2024b)), which mitigates hallucinations by introducing blurred or ambiguous images during decoding to ex-

[*]These authors contributed equally.
[†]Corresponding author.

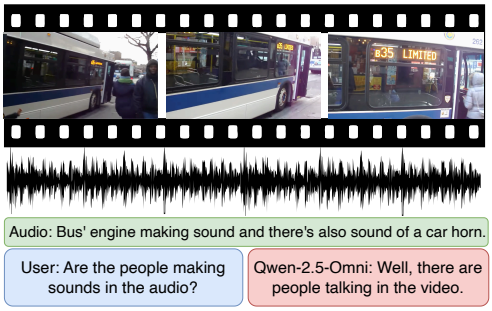


Figure 1: An example where Qwen2.5-Omni-7B hallucinates human speech from visual cues despite audio containing only mechanical noise.

pose and counteract biased language priors; and inference-time intervention (e.g., ICT (Chen et al. 2024b)), which enhances model reliability by injecting intervention vectors into activation layers during the forward pass. However, these methods may inadvertently eliminate all language priors, including those beneficial for reasoning, which can negatively affect performance on reasoning-intensive tasks. Moreover, decoding-based approaches typically require multiple passes during inference, often resulting in significantly increased latency. 2) Training Approaches, which fine-tune models using either synthetic or manually annotated high-quality data to guide the model toward paying more attention to visual information and thus reducing hallucinations. However, manually annotating such data is often prohibitively expensive, while existing synthetic datasets either focus solely on the text modality or fail to capture the intricate relationships between video and its associated audio in omni-modal scenarios.

To address the challenges of overly dominant textual priors and insufficient audio–visual alignment in omni-modal scenarios, we propose OMNIDPO, a preference-based alignment framework that extends DPO to mitigate hallucinations across video, audio, and text modalities. We construct a new omni-modal preference dataset, OMNIDPO-10k, where each sample contains a question with paired textual and audio/video modality preferences (Figure 2). Based on MSR-VTT (Xu et al. 2016) for positive textual preferences, we first use Qwen2-Audio-7B (Chu et al. 2024) to extract audio descriptions, which are then combined with video inputs and fed into Qwen2.5-VL-7B (Bai et al. 2025a) to generate audio-aware answers. Negative samples are produced by inputting the same videos without audio into Qwen2.5-VL-7B, resulting in answers that ignore audio cues. These textual preference pairs explicitly target audio–visual misalignment. To further reduce reliance on textual priors, we introduce noisy video and audio inputs to form modality preference pairs, encouraging attention to visual and auditory signals. Built on OMNIDPO-10k, OMNIDPO extends DPO to omni-modal settings via conditional modality-wise preference learning and incorporates modality-aware losses to explicitly promote grounding in visual and auditory evidence.

Our experimental results demonstrate that applying OMNIDPO to both Qwen2.5-Omni-7B (Xu et al. 2025) and MiniCPM-o-2.6 (Yao et al. 2024) leads to an average performance improvement of 3.48% on the CMM benchmark (Leng et al. 2024a) and 4.23% on AVHBench (Sung-Bin et al. 2025). Notably, beyond effectively reducing omni-modal hallucinations, OMNIDPO also enhances the model's reasoning and QA capabilities in single-modality scenarios. Our contributions can be summarized as:

- We propose OMNIDPO, a direct preference optimization framework tailored for video-audio-text alignment, which addresses the challenge of hallucinations in omni-modal settings by introducing modality-specific conditional preference learning.
- We construct a 10k-sample omni-modal preference dataset OMNIDPO-10k covering diverse real-world scenarios, designed to expose and counteract modality-specific and cross-modal hallucinations. To our best knowledge, OMNIDPO and OMNIDPO-10k are the first hallucination mitigation method and corresponding preference optimization dataset specifically designed for omni-modal scenarios.
- Extensive experiments on Qwen2.5-Omni-7B and MiniCPM-o-2.6 demonstrate that OMNIDPO significantly mitigates hallucinations in omni-modal settings while also enhancing the models' unimodal reasoning capabilities.

## Related Work

### Omni-Modal Large Language Models

Studies on OLLMs aim to provide comprehensive sensory capabilities to large language models(LLMs). These models are designed to process a variety of inputs, such as text, images, videos, and audio, while producing outputs in text or possibly audio formats as well (Yao et al. 2024; Xu et al. 2025; Zhong et al. 2024). OLLMs usually utilize LLMs as their core language models, enhanced with modality-specific tokenizers and encoders to map multimodal inputs into a shared representation space. The powerful language backbones then process these representations to generate outputs based on multimodal inputs and textual instructions (Yin et al. 2024). A significant challenge in this area is aligning the embeddings from different modalities into a unified representation space. Many methods employ text-based or joint text–image embedding spaces, while others anchor embeddings directly in the image space (Girdhar et al. 2023) or treat all modalities equally (Wang et al. 2024d). Progressive alignment strategies have also been proposed to prevent forgetting and enhance cross-modal interactions (Han et al. 2025; Zhang et al. 2024b; Wang et al. 2024c).

### Hallucination in OLLMs

While hallucinations in vision–language models are well studied, omni-modal models remain underexplored. Common causes include modality gaps (Liang et al. 2022; Liu et al. 2024), language priors (Huang et al. 2024a), and statistical biases (Agarwal et al. 2020). Mitigation strategies involve enhanced training with new datasets (Liu et al. 2023;

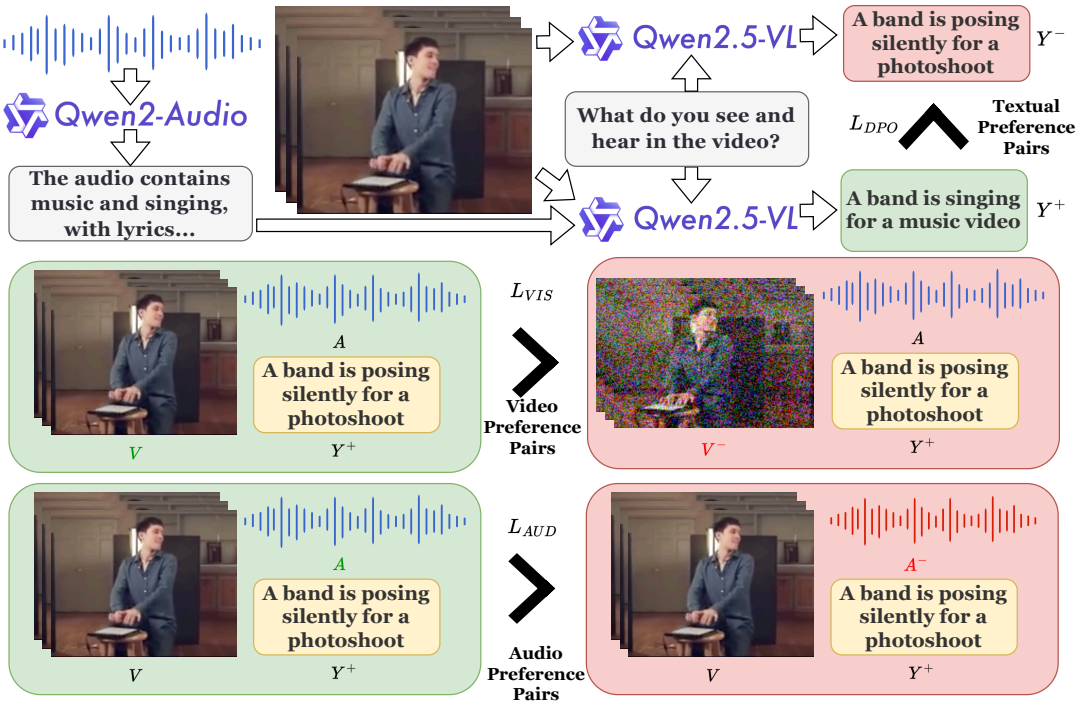


Figure 2: Overview of OMNIDPO.

Gunasekar et al. 2023; Chen et al. 2025; Gunjal, Yin, and Bas 2024; Sun et al. 2023), refined objectives (Chen et al. 2024a; Zhou et al. 2024; Zhao et al. 2024b) and modal-inequality remedies (Wang et al. 2024a; Xie et al. 2024; Xiao et al. 2024; Wu et al. 2024). The second approach addresses inference-time mitigation, including contrastive decoding with logits from perturbed inputs (Chen et al. 2024c; Leng et al. 2023), self-generated intermediate representations (Chuang et al. 2024; Woo et al. 2024; Huo et al. 2025; Li et al. 2025a), and other contrast terms (W. et al. 2024; S. et al. 2024; Y. et al. 2024); on-the-fly hallucination detection and correction (Kuhn, Gal, and Farquhar 2023; Farquhar et al. 2024; Nikitin et al. 2024); image-attention weight manipulation (Zhu et al. 2024; Zhang et al. 2024a; Huo et al. 2024; An et al. 2024; Liu, Zheng, and Chen 2024); prompt-based interventions (Xu et al. 2024; Z. et al. 2024; Wang et al. 2024b); external tool integration (Yin et al. 2023; Zhao et al. 2024a); inference-time steering (Chen et al. 2024b; Li et al. 2025b); and other tactics (Zou et al. 2024).

However, hallucinations in omni-modal LLMs remain an underexplored area, with a few notable exceptions. Research by (Leng et al. 2024a) and (Sung-Bin et al. 2025) has identified modality imbalance as a contributing factor to hallucinations in omni-modal models. Both works developed benchmarks to address this issue, but have not yet provided effective mitigation strategies.

## Dataset Construction

Multimodal hallucinations often arise because text priors dominate other modalities: even when a model receives visual and auditory inputs, it may ignore those signals in favor of familiar textual patterns. Moreover, existing omni-modal training aligns each modality with text independently, overlooking the critical interaction between video and its native audio track. As a result, when understanding hinges on subtle sound cues embedded in the footage, the model can confidently produce incorrect "hallucinated" answers.

To address these issues, we construct two types of preference pairs that explicitly reward audio–visual reasoning:

**Audio–video alignment preferences** To ensure our model truly leverages both audio and video, we first filter out any clip without an audio track, retaining only $\mathcal{D}_0 = \{(V_i, A_i) \mid A_i \neq \emptyset\}$ from MSR-VTT. Each remaining sample is represented as $X = \{V, A, T\}$, where $V$ is the video frames, $A$ the raw audio waveform, and $T$ any accompanying text prompt or question. We then call Qwen2-Audio-7B, denoted $t_a = \mathcal{F}_{\text{audio}}(A)$, which produces a concise textual summary of audio information. Feeding $\{V, t_a\}$ into Qwen2.5-VL-7B (denoted $\mathcal{F}_{\text{VL}}$) yields a multimodal answer $Y^+ = \mathcal{F}_{\text{VL}}(V, t_a)$, whereas masking out the audio text gives $Y^- = \mathcal{F}_{\text{VL}}(V, \varnothing)$. The resulting pair $\{(X, Y^+), (X, Y^-)\}$ teaches the model that genuine audio cues in $X$ should be preferred over video-only reasoning.

**Modality-robustness preferences** Even with aligned audio–video pairs, OLLMs tend to fall back on strong text priors. To resolve this, we create two degraded versions of $X$:

$$X_{V^-} = \{V^-, A, T\}, \quad X_{A^-} = \{V, A^-, T\},$$

where

$$V^- = V + \varepsilon_v, \quad \varepsilon_v \sim \mathcal{N}(0, \sigma_v^2 I), \tag{1}$$

$$A^- = A + \varepsilon_a, \quad \varepsilon_a \sim \mathcal{N}(0, \sigma_a^2 I). \tag{2}$$

| Model | Audio-driven Video Hallucination | | | | | Video-driven Audio Hallucination | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Acc. (↑) | Prec. (↑) | Rec. (↑) | F1 (↑) | Yes (%) | Acc. (↑) | Prec. (↑) | Rec. (↑) | F1 (↑) | Yes (%) |
| Qwen2.5-Omni | | | | | | | | | | |
| Qwen2.5-Omni | 74.12 | 68.72 | **88.56** | 77.38 | 64.44 | 67.60 | 60.84 | **98.78** | 75.30 | 81.18 |
| + VCD | 77.40 | 75.94 | 80.20 | 77.52 | 60.02 | 68.11 | 61.40 | 97.52 | 77.38 | 79.90 |
| + ICT | 76.70 | 73.12 | 84.45 | 78.58 | 61.28 | 68.80 | 62.10 | 96.50 | 77.80 | 78.95 |
| + DPO | 71.74 | 67.73 | 83.04 | 74.61 | 58.19 | 68.70 | 62.41 | 94.04 | 75.03 | 71.72 |
| OMNIDPO | **84.42** | **88.87** | 78.70 | **83.47** | 44.28 | **77.51** | **70.40** | 94.93 | **80.85** | 67.42 |
| MiniCPM-o-2.6 | | | | | | | | | | |
| MiniCPM-o-2.6 | 74.65 | 72.78 | 78.75 | 75.71 | 53.79 | 72.80 | 70.42 | 78.63 | 74.41 | 55.55 |
| + VCD | 73.97 | 72.94 | 76.20 | 74.55 | 53.21 | 73.96 | 72.15 | 78.05 | 75.04 | 55.05 |
| + ICT | 75.92 | 74.83 | 78.12 | 76.46 | 53.32 | 72.25 | 69.88 | 78.20 | 73.92 | 55.18 |
| + DPO | 76.07 | 74.66 | 78.92 | 76.76 | 52.55 | 74.73 | 72.81 | 78.93 | 75.81 | 54.54 |
| OMNIDPO | **76.85** | **74.72** | **81.16** | **77.81** | 54.31 | **76.68** | **74.70** | **80.70** | **77.58** | 54.02 |

Table 1: AVHBench results for Qwen2.5-Omni and MiniCPM-o-2.6 on the two subsets—Audio-driven Video Hallucination and Video-driven Audio Hallucination. Metrics include Accuracy (Acc.), Precision (Prec.), Recall (Rec.), F1-score, and the proportion of "Yes" responses (Yes %).

**Combined dataset** Putting both strategies together, our final dataset is

$$\mathcal{D} = \big\{(X, Y^+), (X, Y^-)\big\} \cup \big\{(X, Y^+), (X_{V^-}, Y^+)\big\} \cup \big\{(X, Y^+), (X_{A^-}, Y^+)\big\}.$$

Providing both audio-video alignment pairs and modality-robustness pairs to drive preference optimization.

To mitigate potential hallucinations from Qwen2-Audio-7B and Qwen2.5-VL-7B, each synthesized sample was independently reviewed by two annotators in a cross-validation procedure, yielding a Fleiss' Kappa of 0.82. Any samples flagged as erroneous or of low quality were subsequently excluded, ensuring the reliability of the final dataset.

In total, we collected 9141 pairs of samples on 1076 distinct videos spanning 20 common categories featuring an open-vocabulary caption task. More information about our dataset found in Appendix C.

## Method

In this section, we detail the OMNIDPO framework for omni-modal hallucination mitigation. We first formalize the problem and recap the DPO objective. We then introduce our conditional multi-modal preference optimization, describing how we generate modality-conditioned preference pairs and incorporate them into the training loss.

### Task Formulation

We consider a general omni-modal input $X = \{V, A, T\}$, where $V$ is visual input (a sequence of video frames, which can degenerate to a single image), $A$ is auditory input (e.g. an audio track or spoken question), and $T$ is textual input (such as a text prompt or question). The OLLM is tasked with generating a textual output $Y$ (an answer or description). Hallucination occurs when $Y$ contains information not supported by $V$ or $A$. We assume access to a dataset of preference comparisons $\{(X, Y^+, Y^-)\}$, where for each input $X$ two candidate outputs are given: $Y^+$ (the preferred/correct output) and $Y^-$ (the less preferred, possibly hallucinated output). These could be obtained via human annotation or automated generation, plus human verification.

DPO provides a way to train the model $P_\theta(Y|X)$ to align with preferences without explicit reinforcement learning. Given a preference tuple $(X, Y^+, Y^-)$, the standard DPO objective encourages the model to increase the probability of $Y^+$ and decrease that of $Y^-$. The DPO loss is:

$$L_{\text{DPO}}(\theta) = -\mathbb{E}_{(X,Y^+,Y^-)}\Big[\log\sigma\Big(\beta\log\Big(\frac{P_\theta(Y^+|X)}{P_{\text{ref}}(Y^+|X)} - \log\frac{P_\theta(Y^-|X)}{P_{\text{ref}}(Y^-|X)}\Big)\Big)\Big], \tag{3}$$

where $\sigma$ is the sigmoid function and $\beta$ is temperature for the preference difference.

### OMNIDPO

As described in the previous section, in addition to the original human preference comparisons between output pairs $(Y^+, Y^-)$ given the full input $X$, we introduce new comparisons using modified inputs $X_{V^-}$ and $X_{A^-}$, which are constructed by masking visual or audio modalities, respectively. We denote the model's output probabilities as $P_\theta(Y|X)$ when full input is given, and $P_\theta(Y|X_{V^-})$, $P_\theta(Y|X_{A^-})$ for the degraded cases.

**Visual Preference Objective:** We create a comparison between the same output under full vs. degraded visual input. Ideally, the model should assign a higher likelihood to the correct output $Y^+$ when it has the full video $V$ than when it only has $V^-$. To enforce this, we treat $(X, Y^+)$ as the preferred scenario and $(X_{V^-}, Y^+)$ as the dispreferred scenario. We do not alter the output in this pair — only the input dif-

fers. The visual preference loss $L_{\text{vis}}$ can be written as:

$$L_{\text{vis}}(\theta) = -\mathbb{E}_{(X,X_{V^-},Y^+)}\Big[\log\sigma\big(\beta\big[\log\frac{P_\theta(Y^+ \mid V,A,T)}{P_{\text{ref}}(Y^+ \mid V,A,T)} - \log\frac{P_\theta(Y^+ \mid V^-,A,T)}{P_{\text{ref}}(Y^+ \mid V^-,A,T)}\big]\big)\Big]. \tag{4}$$

This term pushes the model to increase $P_\theta(Y^+|V,A,T)$ relative to $P_\theta(Y^+|V^-,A,T)$. Intuitively, the model is penalized if it would be just as confident in $Y^+$ even with a missing/blurred video. The only way for the model to satisfy this objective is to genuinely utilize the visual input $V$ when available (since with $V$ it should produce higher confidence in the correct answer). Notably, we do not include $Y^-$ in this comparison; we are not directly contrasting $Y^+$ vs $Y^-$ here, but rather $Y^+$ with vs. without vision.

**Auditory Preference Objective:** Analogously, we define an audio preference loss $L_{\text{aud}}$. We compare the model's confidence in $Y^+$ with full input vs. with muted audio. The audio preference loss is:

$$L_{\text{aud}}(\theta) = -\mathbb{E}_{(X,X_{A^-},Y^+)}\Big[\log\sigma\big(\beta\big[\log\frac{P_\theta(Y^+ \mid V,A,T)}{P_{\text{ref}}(Y^+ \mid V,A,T)} - \log\frac{P_\theta(Y^+ \mid V,A^-,T)}{P_{\text{ref}}(Y^+ \mid V,A^-,T)}\big]\big)\Big]. \tag{5}$$

This term encourages the model to rely on auditory information $A$ when it is present. For example, consider a question $T$ such as "Is the person in the video speaking?", with the correct answer $Y^+$ being "Yes, you can hear them speak." When audio is present, the model should assign high probability to $Y^+$; however, when the audio is removed, its confidence in the same answer should decrease . The loss $L_{\text{aud}}$ enforces this behavior by penalizing overly confident predictions when the relevant auditory evidence is absent. Without such training, a model may answer "Yes" based solely on prior knowledge (e.g., that people in videos often speak), leading to hallucinations when the audio is silent. By explicitly discouraging high-confidence predictions under missing evidence, OMNIDPO mitigates this failure mode. Importantly, degraded-input comparisons are always performed with respect to the preferred output $Y^+$. We assume $Y^+$ is a human-verified answer that is correct given the full input. Training on $(X_{V^-}, Y^+)$ and $(X_{A^-}, Y^+)$ does not assert that $Y^+$ is incorrect for the degraded input; rather, it enforces lower confidence under reduced information. In practice, if $Y^+$ depends on a removed modality (e.g., the question is unanswerable without video or audio), the optimization implicitly encourages the model to assign lower probability to $Y^+$, aligning with the principle that the model should not respond confidently without sufficient perceptual evidence.

**Combining Objectives:** The full OMNIDPO loss combines the standard preference loss with the new modality-specific terms, where $\lambda_V$ and $\lambda_A$ are hyperparameters:

$$L_{\text{OMNI}}(\theta) = L_{\text{DPO}}(\theta) + \lambda_V L_{\text{vis}}(\theta) + \lambda_A L_{\text{aud}}(\theta). \tag{6}$$

# Experiments

## Experiments Setup

**Datasets and Metrics.** **AVHBench** (Sung-Bin et al. 2025) is specifically designed to assess the perception, reasoning, and hallucination robustness of OLLMs. It comprises four evaluation subsets: Audio-Visual Matching, Audio-Visual Captioning, Audio-Driven Video Hallucination, and Video-Driven Audio Hallucination. In our evaluation, we focus on the two hallucination-focused subsets, which specifically measure a model's tendency to hallucinate visual content based solely on audio priors, or vice versa. To quantify performance in hallucination mitigation, the benchmark employs standard classification metrics: Accuracy, Precision, Recall, and F1 Score.

**CMM** (Curse of Multi-Modalities) (Leng et al. 2024a)is a benchmark specifically designed to evaluate hallucination behavior in OLLMs across text, vision, and audio modalities. It focuses on measuring a model's tendency to generate modality-inconsistent outputs by presenting it with inputs where one or more modalities contradict the others. CMM uses two core metrics: Perception Accuracy (PA) and Hallucination Resistance (HR), defined as follows:

$$\begin{aligned}\text{PA} &= \frac{\#\text{ correctly predicted ``yes''}}{\#\text{ ground-truth ``yes''}},\\ \text{HR} &= \frac{\#\text{ correctly predicted ``no''}}{\#\text{ ground-truth ``no''}},\end{aligned} \tag{7}$$

where PA measures the model's ability to correctly detect real elements, while HR quantifies its robustness in rejecting non-existent ones.

**Baselines and Models.** We conduct experiments using two recently released and widely adopted OLLMs: Qwen2.5-Omni and MiniCPM-o-2.6. As baselines, we include both training-free methods, such as VCD and ICT, as well as a training-based method, DPO (text only), which performs preference optimization solely on textual responses.

**Implementation Details.** For both models, we adopt full-parameter fine-tuning using fp16 precision. The learning rate scheduler is set to cosine, with a warmup ratio of 0.1. The DPO loss coefficient ($\beta$) is fixed at 0.1 throughout all training runs. For Qwen2.5-Omni, we use a learning rate of 1e-6, while for MiniCPM-o-2.6, the learning rate is set to 1e-5. We set $\lambda_V = \lambda_A = 1$. All experiments were conducted on a system equipped with 8 × H100 GPUs.

## Main Results

**AVHBench** Table 1 presents the results of Qwen2.5-Omni and MiniCPM-o-2.6 on the two domains of AVHBench: Audio-Driven Video Hallucination and Video-Driven Audio Hallucination. From the results, we draw the following conclusions: **1)** Applying OMNIDPO leads to significant improvements in F1-score, with average gains of 5.82% for Qwen2.5-Omni and 2.64% for MiniCPM-o-2.6. This demonstrates that OMNIDPO, by leveraging both textual preference optimization and omni-modal preference alignment, effectively mitigates hallucination in multimodal settings. **2)** Existing methods such as VCD and ICT, which

| Model | Spurious Inter-modality Correlation | | | | | | Uni-modality Overreliance | | | | | | Overall | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 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 | | | | | | | | | | | | | | |
| Qwen2.5-Omni | 92.0 | 86.5 | **92.0** | 78.0 | 93.0 | 90.5 | 95.0 | 56.5 | 92.5 | 39.5 | 85.5 | 74.0 | 91.7 | 70.2 |
| +VCD | 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 |
| +ICT | **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 |
| +DPO | 92.5 | 87.0 | 92.0 | 79.5 | 95.0 | 93.5 | **95.5** | **66.0** | 94.5 | 37.5 | 87.0 | **76.0** | 92.8 | 73.3 |
| +OMNIDPO | 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** |
| MiniCPM-o-2.6 | | | | | | | | | | | | | | |
| MiniCPM-o-2.6 | 85.0 | 91.0 | 95.0 | 53.0 | 92.0 | 76.5 | 91.0 | 56.5 | 89.0 | 34.5 | 76.5 | 72.0 | 88.1 | 63.9 |
| +VCD | 88.0 | 91.5 | 95.0 | 52.0 | 91.0 | 76.5 | 89.5 | 55.0 | 89.0 | 33.0 | 78.5 | 74.5 | 88.5 | 63.8 |
| +ICT | 89.0 | 93.5 | 94.5 | 54.5 | 93.0 | 76.5 | 88.5 | 56.0 | 90.5 | **35.0** | 78.0 | 75.5 | 88.9 | 65.1 |
| +DPO | 87.0 | 93.0 | 95.5 | 57.5 | **94.5** | 78.5 | 90.5 | 58.5 | **91.5** | 32.5 | **81.5** | 77.0 | 90.1 | 66.2 |
| OMNIDPO | **90.0** | **94.0** | **96.5** | **59.5** | 93.0 | **81.5** | **94.0** | **63.5** | **91.5** | 32.5 | 81.0 | **82.0** | **91.0** | **68.8** |

Table 2: Performance of Qwen2.5-Omni and MiniCPM-o-2.6 on the CMM benchmark after applying different hallucination-mitigation strategies. The benchmark consists of six sub-domains: VL (Visual–Language), AL (Audio–Language), VAL (Visual–Audio–Language), Visual Dom (Visual-dominance), Audio Dom (Audio-dominance), and Lang Dom (Language-dominance). The best results are highlighted in bold.

were originally proposed for vision-language hallucination mitigation, provide only marginal improvements—and in some cases even degrade performance. While these methods reduce the influence of biased language priors and encourage visual grounding, they fail to model the subtle interactions between audio and visual modalities. Moreover, intervention-based approaches like ICT rely on activation shifts observed in visual tasks, which do not generalize well to omni-modal contexts. **3)** For Qwen2.5-Omni, the base model exhibits a strong bias toward answering "yes," often relying on a single modality without properly integrating cross-modal information. This results in hallucinations when critical multimodal cues are absent. OMNIDPO addresses this by constructing multimodal preference pairs that teach the model to be cautious in its affirmations. As a result, the "yes" response rate drops by 16.96%, indicating improved resistance to hallucination in omni-modal scenarios. **4)** While text-only DPO helps the model better understand fine-grained audio-video relationships, it fails to sufficiently shift the model's attention toward non-textual modalities. The dominance of the language modality remains unbalanced, limiting its effectiveness and resulting in no significant performance gains.

**CMM** Table 2 reports the performance of Qwen2.5-Omni and MiniCPM-o-2.6 on the CMM benchmark, which measures hallucination robustness across Spurious Inter-modality Correlation and Uni-modality Overreliance. We draw three key observations. **1)** OMNIDPO consistently improves PA and HR across all sub-domains. Qwen2.5-Omni achieves average gains of +1.6% (pa) and +4.5% (hr), while MiniCPM-o-2.6 improves by +1.9% (pa) and +4.9% (hr). The larger hr gains indicate that OMNIDPO effectively suppresses false affirmations of missing multimodal evidence, consistent with its modality-aware preference training that reduces overconfident hallucinations under insufficient visual or auditory inputs. **2)** Vision-centric methods such as VCD and ICT improve VL performance but degrade other domains, particularly Visual Dominance, suggesting overcorrection toward visual features and weakened cross-modal balance. Moreover, as they do not explicitly model audio-visual interactions, they fail to improve VAL, highlighting their limitations in omni-modal settings. **3)** Standard DPO yields slight VAL improvements from textual preferences but shows minimal gains or regressions elsewhere, notably in Audio and Visual Dominance. In contrast, OMNIDPO improves all sub-domains by explicitly encouraging discrimination between complete and degraded modality inputs.

## Analysis

### Ablation Study

To investigate how preference optimization for different modalities affects a model's resistance to hallucination, we conduct an ablation study on Qwen2.5-Omni-7B. The results are presented in Table 3. From the table, we observe that applying either video-only or audio-only preference optimization individually yields positive effects on the model's performance. When both modalities are optimized simultaneously, the model's robustness against hallucination is further enhanced. Interestingly, on the CMM benchmark, we notice a modality interference effect: optimizing preferences for only one modality can negatively impact the performance related to the other modality. In contrast, jointly optimizing both audio and video preferences leads to consistent and overall improvements across modalities.

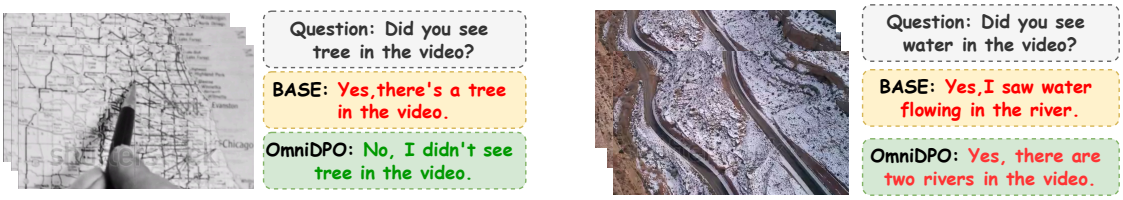


Figure 3: Case Study and Error Analysis of OMNIDPO.

## Experiments on Non-hallucination Benchmarks

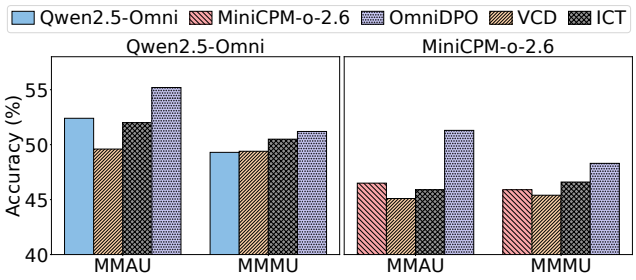


Figure 4: Performance comparison on reasoning benchmarks (MMAU (Sakshi et al. 2024) and MMMU (Yue et al. 2024)) after applying different alignment methods. All evaluations are performed in a zero-shot setting.

To investigate whether applying OMNIDPO affects a model's general reasoning capabilities, we evaluate performance on two widely used benchmarks: MMAU (Massive Multi-Task Audio Understanding) (Sakshi et al. 2024) and MMMU (Massive Multi-discipline Multimodal Understanding) (Yue et al. 2024). We also compare our results with those of VCD and ICT. The outcomes are shown in Figure 4. Experimental results show that after applying OMNIDPO, Qwen2.5-Omni-7b achieves an average performance gain of 2.4%, while MiniCPM-o-2.6 improves by 3.6%. This indicates that OMNIDPO not only mitigates hallucinations but also enhances the model's reasoning and understanding capabilities by encouraging better attention to audio and visual inputs through multimodal preference optimization. In contrast, VCD reduces performance on reasoning tasks, likely because it removes language priors, including those beneficial for reasoning. ICT, while improving performance on MMMU by reinforcing visual grounding, does not account for audio modality and thus fails to improve the model's understanding of auditory information.

## Case Study and Error Analysis

In Figure 3, we present a case study from the CMM benchmark to evaluate the performance of our model, OMNIDPO. As illustrated on the left side of the figure, while the base model, Qwen2.5-Omni-7B, failed to distinguish between maps with trees, our OMNIDPO completed the task. In this instance, OMNIDPO overcame the tendency to answer "Yes" due to language priors and instead derived the correct response from the visual information.

| Model | CMM | | | | | | AVHBench | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | VL | | AL | | VAL | | A-V Hal. | | V-A Hal. | |
| | pa ↑ | hr ↑ | pa ↑ | hr ↑ | pa ↑ | hr ↑ | Acc. ↑ | F1 ↑ | Acc. ↑ | F1 ↑ |
| Qwen2.5-Omni | 92.0 | 86.5 | 92.0 | 78.0 | 93.0 | 90.5 | 74.1 | 77.4 | 67.6 | 75.3 |
| + Audio | 93.5 | 82.5 | **93.0** | 82.5 | 92.5 | 93.5 | 80.9 | 81.1 | 76.7 | 80.4 |
| + Video | **95.5** | 88.5 | 90.5 | 63.0 | 94.0 | 66.0 | 82.4 | 81.4 | 66.3 | 74.6 |
| + OMNIDPO | 94.0 | **89.5** | 91.5 | **83.5** | **95.5** | **94.5** | **84.4** | **83.5** | **77.5** | **80.9** |

Table 3: Ablation Study. "+ Audio" denotes applying only the audio loss $L_{\text{aud}}$, while "+ Video" uses only the visual loss $L_{\text{vis}}$. "A-V Hal." refers to the Audio-driven Video Hallucination setting, and "V-A Hal." to the Video-driven Audio Hallucination setting.

On the right side of the figure, however, both the base model and OMNIDPO produced incorrect answers. We hypothesize that this is due to limitations in the base model's visual understanding. Although OMNIDPO mitigates the issue of over-reliance on textual modality, it does not fundamentally enhance the visual capabilities inherited from the base model. Consequently, when faced with ambiguous visual content, such as in the right part of Figure 3, where roads closely resemble rivers, the model still struggles to provide an accurate response.

## Conclusion and Limitations

We presented OMNIDPO, the first dedicated framework for mitigating omni-modal hallucinations in OLLMs. By extending DPO with modality-specific objectives, OMNIDPO compels models to ground outputs in both visual and auditory inputs, rather than over-relying on textual priors or spurious correlations. To support training, we constructed OmniDPO-10k, the first preference-alignment dataset designed to expose and counteract hallucinations from misaligned or over-relied modalities (text, video, audio). Each sample includes paired modality-based preferences, enabling fine-grained supervision of model behavior. Through experiments on state-of-the-art benchmarks and models of different scales, we demonstrated that OMNIDPO substantially reduces hallucinations across video, audio, and text scenarios, outperforming existing alignment methods.

**Limitations** Due to limitations in the base models, our work currently supports only textual, visual, and audio modalities. Future research may explore the incorporation of more diverse input forms.

## Acknowledgements

This work is supported by the National Key Research and Development Program of China (No.2024YFB3309702), the National Nature Science Foundation of China (No.62021002), Tsinghua BNRist and Beijing Key Laboratory of Industrial Big Data System and Application, Beijing Natural Science Foundation(QY25044).

## References

Agarwal; et al. 2020. Towards Causal VQA: Revealing and Reducing Spurious Correlations by Invariant and Covariant Semantic Editing. arXiv:1912.07538.

An, W.; Tian, F.; Leng, S.; Nie, J.; Lin, H.; Wang, Q.; Dai, G.; Chen, P.; and Lu, S. 2024. AGLA: Mitigating Object Hallucinations in Large Vision-Language Models with Assembly of Global and Local Attention. *arXiv preprint*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025a. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Bai, Z.; Wang, P.; Xiao, T.; He, T.; Han, Z.; Zhang, Z.; and Shou, M. Z. 2025b. Hallucination of Multimodal Large Language Models: A Survey. arXiv:2404.18930.

Chen, B.; Lyu, X.; Gao, L.; Song, J.; and Shen, H. T. 2024a. Alleviating Hallucinations in Large Vision-Language Models through Hallucination-Induced Optimization. arXiv:2405.15356.

Chen, C.; Liu, M.; Jing, C.; Zhou, Y.; Rao, F.; Chen, H.; Zhang, B.; and Shen, C. 2025. PerturboLLaVA: Reducing Multimodal Hallucinations with Perturbative Visual Training. arXiv:2503.06486.

Chen, J.; Zhang, T.; Huang, S.; Niu, Y.; Zhang, L.; Wen, L.; and Hu, X. 2024b. ICT: Image-Object Cross-Level Trusted Intervention for Mitigating Object Hallucination in Large Vision-Language Models. arXiv:2411.15268.

Chen, Z.; Zhao, Z.; Luo, H.; Yao, H.; Li, B.; and Zhou, J. 2024c. HALC: Object Hallucination Reduction via Adaptive Focal-Contrast Decoding. http://arxiv.org/abs/2403.00425, arXiv:2403.00425.

Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; Zhou, C.; and Zhou, J. 2024. Qwen2-Audio Technical Report. arXiv:2407.10759.

Chuang, Y.-S.; Xie, Y.; Luo, H.; Kim, Y.; Glass, J.; and He, P. 2024. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models. http://arxiv.org/abs/2309.03883, arXiv:2309.03883.

Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017): 625–630.

Favero, A.; Zancato, L.; Trager, M.; Choudhary, S.; Perera, P.; Achille, A.; Swaminathan, A.; and Soatto, S. 2024. Multimodal Hallucination Control by Visual Information Grounding. arXiv:2403.14003.

Gao, H.; Qu, J.; Tang, J.; Bi, B.; Liu, Y.; Chen, H.; Liang, L.; Su, L.; and Huang, Q. 2025. Exploring Hallucination of Large Multimodal Models in Video Understanding: Benchmark, Analysis and Mitigation. arXiv:2503.19622.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. ImageBind: One Embedding Space To Bind Them All. arXiv:2305.05665.

Gunasekar, S.; Zhang, Y.; Aneja, J.; Mendes, C. C. T.; Giorno, A. D.; Gopi, S.; Javaheripi, M.; Kauffmann, P.; de Rosa, G.; Saarikivi, O.; Salim, A.; Shah, S.; Behl, H. S.; Wang, X.; Bubeck, S.; Eldan, R.; Kalai, A. T.; Lee, Y. T.; and Li, Y. 2023. Textbooks Are All You Need. arXiv:2306.11644.

Gunjal, A.; Yin, J.; and Bas, E. 2024. Detecting and Preventing Hallucinations in Large Vision Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18135–18143.

Guo, Y.; Ma, S.; Ma, S.; Bao, X.; Xie, C.-W.; Zheng, K.; Weng, T.; Sun, S.; Zheng, Y.; and Zou, W. 2025. Aligned Better, Listen Better for Audio-Visual Large Language Models. arXiv:2504.02061.

Han, J.; Gong, K.; Zhang, Y.; Wang, J.; Zhang, K.; Lin, D.; Qiao, Y.; Gao, P.; and Yue, X. 2025. OneLLM: One Framework to Align All Modalities with Language. arXiv:2312.03700.

Huang, Q.; Dong, X.; Zhang, P.; Wang, B.; He, C.; Wang, J.; Lin, D.; Zhang, W.; and Yu, N. 2024a. OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation. *arXiv preprint*.

Huang, W.; Liu, H.; Guo, M.; and Gong, N. Z. 2024b. Visual Hallucinations of Multi-modal Large Language Models. arXiv:2402.14683.

Huo, F.; Xu, W.; Zhang, Z.; Wang, H.; Chen, Z.; and Zhao, P. 2024. Self-Introspective Decoding: Alleviating Hallucinations for Large Vision-Language Models. arXiv:2408.02032.

Huo, F.; Xu, W.; Zhang, Z.; Wang, H.; Chen, Z.; and Zhao, P. 2025. Self-Introspective Decoding: Alleviating Hallucinations for Large Vision-Language Models. arXiv:2408.02032.

Kuhn, L.; Gal, Y.; and Farquhar, S. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. *arXiv*.

Leng, S.; Xing, Y.; Cheng, Z.; Zhou, Y.; Zhang, H.; Li, X.; Zhao, D.; Lu, S.; Miao, C.; and Bing, L. 2024a. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. *arXiv preprint arXiv:2410.12787*.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2023. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. arXiv:2311.16922.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2024b. 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*, 13872–13882.

Li, J.; Zhang, J.; Jie, Z.; Ma, L.; and Li, G. 2025a. Mitigating Hallucination for Large Vision Language Model by Inter-Modality Correlation Calibration Decoding. arXiv:2501.01926.

Li, Z.; Shi, H.; Gao, Y.; Liu, D.; Wang, Z.; Chen, Y.; Liu, T.; Zhao, L.; Wang, H.; and Metaxas, D. N. 2025b. The Hidden Life of Tokens: Reducing Hallucination of Large Vision-Language Models via Visual Information Steering. arXiv:2502.03628.

Liang, W.; Zhang, Y.; Kwon, Y.; Yeung, S.; and Zou, J. 2022. Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. arXiv:2203.02053.

Liu, F.; Lin, K.; Li, L.; Wang, J.; Yacoob, Y.; and Wang, L. 2023. Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. In *The Twelfth International Conference on Learning Representations*.

Liu, H.; Xue, W.; Chen, Y.; Chen, D.; Zhao, X.; Wang, K.; Hou, L.; Li, R.; and Peng, W. 2024. A Survey on Hallucination in Large Vision-Language Models. arXiv:2402.00253.

Liu, S.; Zheng, K.; and Chen, W. 2024. Paying More Attention to Image: A Training-Free Method for Alleviating Hallucination in LVLMs. arXiv:2407.21771.

Nikitin, A.; Kossen, J.; Gal, Y.; and Marttinen, P. 2024. Kernel Language Entropy: Fine-Grained Uncertainty Quantification for LLMs from Semantic Similarities. *arXiv*.

Nishimura, T.; Nakada, S.; and Kondo, M. 2024. On the Audio Hallucinations in Large Audio-Video Language Models. arXiv:2401.09774.

S., K.; B., C.; S., B.; S., A.; and S.Y., Y. 2024. Vacode: Visual Augmented Contrastive Decoding. *arXiv preprint arXiv:2408.05337*.

Sahoo, P.; Meharia, P.; Ghosh, A.; Saha, S.; Jain, V.; and Chadha, A. 2024. A Comprehensive Survey of Hallucination in Large Language, Image, Video and Audio Foundation Models. arXiv:2405.09589.

Sakshi, S.; Tyagi, U.; Kumar, S.; Seth, A.; Selvakumar, R.; Nieto, O.; Duraiswami, R.; Ghosh, S.; and Manocha, D. 2024. MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark. arXiv:2410.19168.

Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.-Y.; Wang, Y.-X.; Yang, Y.; Keutzer, K.; and Darrell, T. 2023. Aligning Large Multimodal Models with Factually Augmented RLHF. arXiv:2309.14525.

Sung-Bin, K.; Hyun-Bin, O.; Lee, J.; Senocak, A.; Chung, J. S.; and Oh, T.-H. 2025. AVHBench: A Cross-Modal Hallucination Benchmark for Audio-Visual Large Language Models. arXiv:2410.18325.

W., Z.; X., F.; L., Z.; Q., L.; L., H.; Y., G.; W., M.; Y., X.; and B., Q. 2024. Investigating and Mitigating the Multimodal Hallucination Snowballing in Large Vision-Language Models. *arXiv preprint arXiv:2407.00569*.

Wang, F.; Zhou, W.; Huang, J. Y.; Xu, N.; Zhang, S.; Poon, H.; and Chen, M. 2024a. mDPO: Conditional Preference Optimization for Multimodal Large Language Models. arXiv:2406.11839.

Wang, J. 2024. Hallucination Reduction and Optimization for Large Language Model-Based Autonomous Driving. *Symmetry*, 16(9).

Wang, X.; Pan, J.; Ding, L.; and Biemann, C. 2024b. Mitigating Hallucinations in Large Vision-Language Models with Instruction Contrastive Decoding. arXiv:2403.18715.

Wang, Z.; Zhang, Z.; Cheng, X.; Huang, R.; Liu, L.; Ye, Z.; Huang, H.; Zhao, Y.; Jin, T.; Gao, P.; and Zhao, Z. 2024c. FreeBind: Free Lunch in Unified Multimodal Space via Knowledge Fusion. arXiv:2405.04883.

Wang, Z.; Zhang, Z.; Zhang, H.; Liu, L.; Huang, R.; Cheng, X.; Zhao, H.; and Zhao, Z. 2024d. OmniBind: Large-scale Omni Multimodal Representation via Binding Spaces. arXiv:2407.11895.

Woo, S.; Kim, D.; Jang, J.; Choi, Y.; and Kim, C. 2024. Don't Miss the Forest for the Trees: Attentional Vision Calibration for Large Vision Language Models. arXiv:2405.17820.

Wu, K.; Jiang, B.; Jiang, Z.; He, Q.; Luo, D.; Wang, S.; Liu, Q.; and Wang, C. 2024. NoiseBoost: Alleviating Hallucination with Noise Perturbation for Multimodal Large Language Models. http://arxiv.org/abs/2405.20081, arXiv:2405.20081.

Xiao, X.; Wu, B.; Wang, J.; Li, C.; Zhou, X.; and Guo, H. 2024. Seeing the Image: Prioritizing Visual Correlation by Contrastive Alignment. arXiv:2405.17871.

Xie, Y.; Li, G.; Xu, X.; and Kan, M.-Y. 2024. V-DPO: Mitigating Hallucination in Large Vision Language Models via Vision-Guided Direct Preference Optimization. arXiv:2411.02712.

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.; and Lin, J. 2025. Qwen2.5-Omni Technical Report. arXiv:2503.20215.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5288–5296. Las Vegas, NV, USA: IEEE.

Xu, X.; Tao, C.; Shen, T.; Xu, C.; Xu, H.; Long, G.; and guang Lou, J. 2024. Re-Reading Improves Reasoning in Large Language Models. http://arxiv.org/abs/2309.06275, arXiv:2309.06275.

Y., P.; D., L.; J., C.; and B., C. 2024. ConVis: Contrastive Decoding with Hallucination Visualization for Mitigating Hallucinations in Multimodal Large Language Models. *arXiv preprint arXiv:2408.13906*.

Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. MiniCPM-o 2.6: A GPT-4o Level MLLM for Vision, Speech and Multimodal Live Streaming on Your Phone. https://github.com/OpenBMB/MiniCPM-o.

Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; and Chen, E. 2024. A survey on multimodal large language models. *National Science Review*, 11(12).

Yin, S.; Fu, C.; Zhao, S.; Xu, T.; Wang, H.; Sui, D.; Shen, Y.; Li, K.; Sun, X.; and Chen, E. 2023. Woodpecker: Hallucination Correction for Multimodal Large Language Models. *arXiv preprint*.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; Wei, C.; Yu, B.; Yuan, R.; Sun, R.; Yin, M.; Zheng, B.; Yang, Z.; Liu, Y.; Huang, W.; Sun, H.; Su, Y.; and Chen, W. 2024. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of CVPR*.

Z., H.; Z., B.; H., M.; Q., X.; C., Z.; and M.Z., S. 2024. Skip: A Simple Method to Reduce Hallucination in Large Vision-Language Models. *arXiv preprint arXiv:2402.01345*.

Zhang, Y.-F.; Yu, W.; Wen, Q.; Wang, X.; Zhang, Z.; Wang, L.; Jin, R.; and Tan, T. 2024a. Debiasing Multimodal Large Language Models. *arXiv preprint*.

Zhang, Z.; Wang, Z.; Liu, L.; Huang, R.; Cheng, X.; Ye, Z.; Liu, H.; Huang, H.; Zhao, Y.; Jin, T.; et al. 2024b. Extending multimodal contrastive representations. *Advances in Neural Information Processing Systems*, 37: 91880–91903.

Zhao, L.; Deng, Y.; Zhang, W.; and Gu, Q. 2024a. Mitigating object hallucination in large vision-language models via classifier-free guidance.

Zhao, Z.; Wang, B.; Ouyang, L.; Dong, X.; Wang, J.; and He, C. 2024b. Beyond Hallucinations: Enhancing LVLMs through Hallucination-Aware Direct Preference Optimization. arXiv:2311.16839.

Zhong, Z.; Wang, C.; Liu, Y.; Yang, S.; Tang, L.; Zhang, Y.; Li, J.; Qu, T.; Li, Y.; Chen, Y.; Yu, S.; Wu, S.; Lo, E.; Liu, S.; and Jia, J. 2024. Lyra: An Efficient and Speech-Centric Framework for Omni-Cognition. arXiv:2412.09501.

Zhou, Y.; Cui, C.; Rafailov, R.; Finn, C.; and Yao, H. 2024. Aligning Modalities in Vision Large Language Models via Preference Fine-tuning. arXiv:2402.11411.

Zhu, L.; Ji, D.; Chen, T.; Xu, P.; Ye, J.; and Liu, J. 2024. IBD: Alleviating Hallucinations in Large Vision-Language Models via Image-Biased Decoding. *arXiv preprint*.

Zou, X.; Wang, Y.; Yan, Y.; Huang, S.; Zheng, K.; Chen, J.; Tang, C.; and Hu, X. 2024. Look Twice Before You Answer: Memory-Space Visual Retracing for Hallucination Mitigation in Multimodal Large Language Models. arXiv:2410.03577.