# TAG-GRPO: MITIGATING CROSS-MODAL HALLUCINATIONS VIA TASK-AWARE GROUPED GRPO

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Omni-modal large language models jointly process audio, video, and language, yet they often produce correct answers through incorrect modality reasoning—relying on spurious correlations rather than task-relevant evidence, which results in cross-modal hallucinations in audio-visual understanding. Existing modality-aware preference optimization methods mitigate this issue via offline perturbation pairs; however, their static supervision struggles to capture task-dependent modality roles. To address this, we propose **Task-aware Grouped GRPO (TAG-GRPO)**, an online policy optimization framework that extends GRPO from standard response-only rollouts to task-aware, modality-conditioned rollouts. Specifically, TAG-GRPO generates task-aware modality conditions through diffusion-based corruption and routes them into *Intact*, *Robustness*, and *Sensitivity* groups according to the task type. Group-wise advantage normalization then independently optimizes correctness under full evidence, invariance to irrelevant modality perturbations, and sensitivity to the disruption of task-relevant evidence. This explicit behavioral separation encourages the model to ground its reasoning in proper modality evidence, thereby suppressing cross-modal hallucinations beyond what answer-level correctness alone can achieve. Extensive experiments on AVHBench and CMM demonstrate that TAG-GRPO improves audio-visual grounding and reduces cross-modal hallucinations, highlighting the critical importance of optimizing modality-faithful behavior beyond answer correctness.

## 1 INTRODUCTION

Recent omni-modal large language models have advanced beyond vision-language understanding by jointly processing audio, video, and language within unified architectures Xu et al. (2025a); AI et al. (2025); Xu et al. (2025b); Hurst et al. (2024), following earlier video and audio-visual LLMs such as Video-ChatGPT and Video-LLaMA Maaz et al. (2024); Zhang et al. (2023). These models enable a broad range of audio-visual reasoning tasks, but improved multimodal capability does not necessarily guarantee reliable modality grounding. In many cases, models can generate correct answers by relying on spurious correlations, language priors, or dominant modalities rather than the evidence required by the task. Such failures, where the answer is correct but the underlying modal reasoning is incorrect, lead to cross-modal hallucinations in audio-visual understanding.

Recent benchmarks reveal that audio-visual models frequently struggle with modality faithfulness. While general multimodal benchmarks such as MME evaluate broad perception and cognition abilities Fu et al. (2026), hallucination-oriented studies show that high-level capability does not necessarily imply faithful grounding Rohrbach et al. (2018). In particular, AVHBench Sung-Bin et al. (2025) identifies audio-driven visual and video-driven audio hallucinations, indicating that models often mistake confuse cross-modal correlations with actual evidence. CMM Leng et al. (2024a) further demonstrates that these models exhibit modality bias and fail to recognize inconsistent audio-visual relationships. More recently, AVTrustBench and SVHalluc extend this line of analysis to modality reliability and speech-vision hallucination, underscoring that robust audio-visual reasoning hinges on the ability to separate task-relevant evidence from misleading modality cues Chowdhury et al. (2025); Zhang et al. (2026). Thus, hallucination mitigation should optimize not only answer correctness but also the appropriateness of the modality evidence on which the model relies.



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.

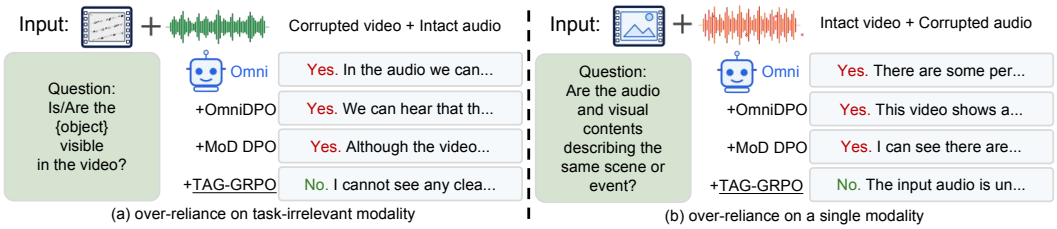



Figure 1: Failure patterns of omni-modal models under task-aware modality corruption. Baseline models may rely on irrelevant audio cues when task-relevant visual evidence is corrupted (a), or infer audio-visual consistency from the intact video alone (b). TAG-GRPO instead responds sensitively to relevant-modality corruption and disrupted cross-modal relations.

Preference optimization offers a promising direction for improving multimodal faithfulness. For example, RLHF and Direct Preference Optimization (DPO) align model behaviors using reward signals or preference comparisons Ouyang et al. (2022); Rafailov et al. (2023). Silkie Li et al. (2023a) suggests that preference supervision in vision-language models can enhance both helpfulness and faithfulness. Building on this idea, multimodal extensions such as HA-DPO, mDPO, and OPA-DPO construct hallucination-aware preference signals to promote grounded responses Zhao et al. (2023); Wang et al. (2024); Yang et al. (2025). OmniDPO and MoD-DPO further introduce modality perturbations, explicitly regulating modality usage by contrasting complete and corrupted inputs Chen et al. (2026); Chaubey et al. (2026). However, these methods rely on offline pairwise preference supervision, where modality relations are predefined before training. Also, task-adaptive methods like MAD and SFFL argue that modality reliance should be adjusted per task rather than fixed globally Chung et al. (2026); Li et al. (2026). Such static supervision cannot fully capture task-dependent modality roles: the same modality perturbation may demand robustness for one task but sensitivity for another. Recently, CFPO Yu et al. (2026) integrates counterfactual interventions into GRPO to enforce causal consistency; however, its policy remains fixed, preventing adaptation to task-dependent modality dynamics, and is restricted to the visual modality.

To this end, we propose **Task-aware Grouped GRPO** (TAG-GRPO), an online policy optimization framework that reformulates modality perturbation learning as task-aware behavioral optimization. Instead of constructing fixed preference pairs, TAG-GRPO performs modality-conditioned counterfactual rollouts and partitions responses into three groups: *Intact*, *Robustness*, and *Sensitivity*. The Intact group optimizes reasoning with complete evidence, the Robustness group encourages invariance to irrelevant modality corruption, and the Sensitivity group encourages adaptation when task-relevant evidence or cross-modal relationships are disrupted. By explicitly separating these behavioral objectives during online optimization, TAG-GRPO enables modality-faithful reasoning beyond answer correctness.

The contributions of this work are summarized as follows:

- We propose **TAG-GRPO**, a task-aware grouped policy optimization framework that extends GRPO to modality-conditioned counterfactual rollouts for improving audio-visual faithfulness.
- We introduce a behavior-oriented supervision paradigm separating intact, robustness, and sensitivity objectives for online optimization of task-dependent modality behaviors.
- Extensive experiments demonstrate that optimizing modality-faithful behaviors beyond answer correctness reduces cross-modal hallucinations and improves audio-visual grounding.

## 2 Related Work

**Cross-modal Hallucination in Audio-Visual Models.** Hallucination evaluation has expanded from vision-language models (POPE, MMHal-Bench, HallusionBench) Li et al. (2023b); Sun et al. (2024); Guan et al. (2024) to audio-visual settings. Benchmarks such as AVHBench, CMM, AVTrustBench, and SVHalluc Sung-Bin et al. (2025); Leng et al. (2024a); Chowdhury et al. (2025); Zhang et al. (2026) demonstrate that models frequently rely on spurious cross-modal correlations rather than task-relevant evidence, resulting in modality grounding failures. Analyses of modality interference and bias Cai et al. (2025); Zheng et al. (2025) further attribute these issues to optimization objectives

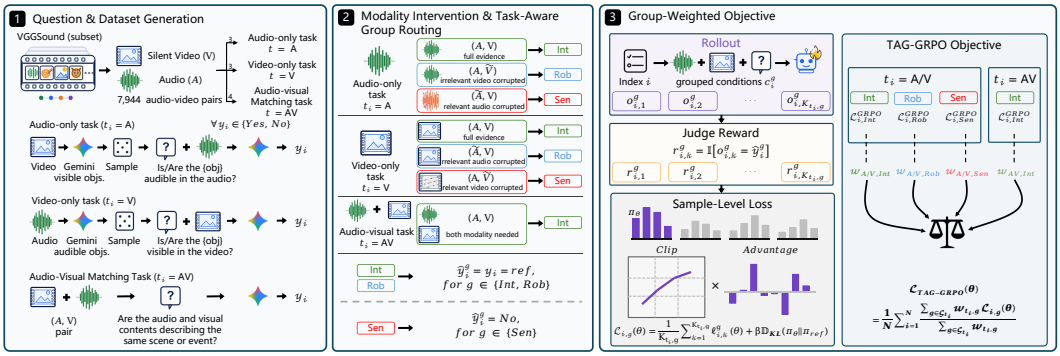



Figure 2: Overview of TAG-GRPO. Given audio–visual pairs from VGGSound Chen et al. (2020), TAG-GRPO constructs modality-specific and cross-modal tasks, applies modality interventions, and routes samples into Intact, Robustness, and Sensitivity groups. Group-wise rollouts and rewards are aggregated with normalized advantages to optimize the task- and group-weighted GRPO objective.

and data imbalance, motivating training strategies that explicitly regulate modality usage beyond answer correctness.

**Modality-aware Preference Optimization.** Preference optimization (e.g., RLHF and DPO) Ouyang et al. (2022); Rafailov et al. (2023) has been adapted to multimodal faithfulness through hallucination-aware preference signals (HA-DPO, mDPO, OPA-DPO) Zhao et al. (2023); Wang et al. (2024); Yang et al. (2025) and modality perturbation (OmniDPO, MoD-DPO) Chen et al. (2026); Chaubey et al. (2026). These methods encourage models to rely on authentic multimodal evidence; however, they depend on offline pairwise preference construction with static modality relationships, which provides limited flexibility for tasks that require distinct modality behaviors under varying conditions.

**Online Reinforcement Learning for Multimodal Alignment.** Online policy optimization, exemplified by GRPO Shao et al. (2024), enables efficient alignment via grouped rollouts and has shown promise in omni-modal reasoning (Omni-R1) Rouditchenko et al. (2025). Other online or adaptive preference methods (OViP, MBPO) Liu et al. (2025b;a); Yu et al. (2026) target response quality or modality balance, yet do not explicitly optimize task-dependent modality behaviors. In contrast, TAG-GRPO introduces task-aware modality-conditioned grouped rollouts, enabling online optimization of distinct modality behaviors under different evidence conditions.

## 3 Preliminaries

**Task Definition.** In this paper, we study modality-faithful reasoning for omni-modal models with paired audio and video inputs. Formally, each instance is formulated as $(A, V, q, t, y) \sim \mathcal{D}$, where $A$ and $V$ denote the audio and video streams, $q$ is the query, $t \in \{\mathsf{A}, \mathsf{V}, \mathsf{AV}\}$ indicates the task type, and $y \in \{\mathrm{Yes}, \mathrm{No}\}$ is the reference answer. Although omni-modal models receive multiple input modalities, different tasks rely on varying sources of evidence. Specifically, an audio-only task ($t = \mathsf{A}$) should be answered based on acoustic information, a video-only task ($t = \mathsf{V}$) should rely on visual information, and an audio-visual matching task ($t = \mathsf{AV}$) requires jointly reasoning about the correspondence between audio and visual streams. Thus, we define the task-relevant modality sets as: $\mathcal{M}_{\mathsf{A}} = \{A\}, \mathcal{M}_{\mathsf{V}} = \{V\}, \mathcal{M}_{\mathsf{AV}} = \{A, V\}$.

**Diffusion-Based Modality Strategy.** We adopt the diffusion-based corruption strategy from Visual Contrastive Decoding to generate corrupted modality inputs Leng et al. (2024b). Specifically, we define a standard forward diffusion process over $T = 1000$ discrete steps indexed by $s \in \{0, \cdots, T-1\}$. The noise variance at step $s$ follows a sigmoid-shaped schedule $\zeta_s = \sigma\left(-6 + \frac{12s}{T-1}\right)\left(5 \times 10^{-3} - 10^{-5}\right) + 10^{-5}$, where $\sigma(\cdot)$ denotes the sigmoid function. Let $\alpha_s = 1 - \zeta_s$ and $\bar{\alpha}_s = \prod_{k=0}^{s} \alpha_k$. For an original modality input $z$, its corrupted counterpart at timestep $s$ is sampled as:

$$\tilde{z} = \sqrt{\bar{\alpha}_s}\, z + \sqrt{1 - \bar{\alpha}_s}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{1}$$

**Group Relative Policy Optimization.** Group Relative Policy Optimization (GRPO) Shao et al. (2024) is an online policy optimization method that removes the need for an additional value model. Given an input $x$, GRPO samples a group of responses $\{o_i\}_{i=1}^{G}$ from the old policy $\pi_{\theta_{\text{old}}}(\cdot|x)$, and obtains their rewards $\{r_i\}_{i=1}^{G}$. The relative advantage is estimated by normalizing rewards within the sampled group: $\hat{A}_i = \frac{r_i - \mu_r}{\sigma_r + \varepsilon}$, where $\mu_r$ and $\sigma_r$ denote the mean and standard deviation of group rewards, and $\varepsilon$ is a small constant for numerical stability. The policy $\pi_\theta$ is then optimized using the following objective:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\mathcal{L}_i(\theta) + \beta\mathbb{D}_{\text{KL}}(\pi_\theta||\pi_{\text{ref}}), \tag{2}$$

$$\mathcal{L}_i(\theta) = -\frac{1}{M_i}\sum_{m=1}^{M_i}\min\left(\rho_{i,m}(\theta)\hat{A}_i, \text{clip}(\rho_{i,m}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i\right), \tag{3}$$

where $\rho_{i,m}(\theta) = \frac{\pi_\theta(o_{i,m}|x, o_{i,<m})}{\pi_{\theta_{\text{old}}}(o_{i,m}|x, o_{i,<m})}$ is the policy probability ratio of the $m$-th token in the response $o_i$ of length $M_i$. $\mathbb{D}_{\text{KL}}(\pi_\theta||\pi_{\text{ref}})$ is the regularization term used to prevent excessive deviation from the reference policy $\pi_{ref}$, controlled by coefficient $\beta$. $\epsilon$ is used to controlling the update step.

# 4 METHOD

In this section, we introduce the proposed TAG-GRPO, which consists of four components: *Question & Dataset Generation*, *Task-Aware Modality Intervention*, *Task-Aware Group Routing*, and the *Group-Weighted Objective*.

## 4.1 QUESTION & DATASET GENERATION

To instantiate the task taxonomy $\mathsf{A}, \mathsf{V}$, and $\mathsf{AV}$, we construct a curated training set consisting of binary question-answer pairs designed to probe cross-modal hallucinations. Specifically, we source our data from VGGSound Chen et al. (2020); MTEB (2026), a large-scale audio-visual dataset comprising unconstrained videos with ambient soundtracks. To ensure data quality, we discard videos that fail to decode with `ffmpeg` or `decord`, yielding a total of 7,944 valid audio-video pairs. We then partition them across the three task types and generate one task-specific binary question-answer pair. In this way, each retained audio-video pair is associated with a single task type $t \in \{\mathsf{A}, \mathsf{V}, \mathsf{AV}\}$. Moreover, we use Gemini-3.1-Pro-Preview Google (2026) as a high-capacity annotator (abbreviated as Gemini) to obtain reliable reference annotations $y$. Thereafter, we detail the generation process for each task.

Table 1: Distribution of training labels $y_i$ for different task types constructed from VGGSound.

| **Task Type** | **Yes** | **No** | **Total** |
|---|---|---|---|
| Audio-Only ($\mathsf{A}$) | 1,519 | 864 | 2,383 |
| Video-Only ($\mathsf{V}$) | 903 | 1,480 | 2,383 |
| Audio-Visual Matching ($\mathsf{AV}$) | 2,611 | 567 | 3,178 |
| **Total** | 5,033 | 2,911 | 7,944 |

***For the audio-only task***, reference answers must rely solely on acoustic evidence. To generate queries, we feed only the video stream to Gemini using the prompt (Appendix A.1.1) to list visible non-human objects likely to produce characteristic sounds. We randomly sample one object and ask: "*Is/Are the* $\{$*object*$\}$ *audible in the audio? Answer Yes or No.*" This question, together with the intact audio, is then provided to Gemini for the reference answer. This cross-modal design ensures that answers are audio-grounded while candidates derive semantically from visual cues, enabling controlled modality grounding. ***For the video-only task***, we mirror the audio-only design: feed only the audio stream to Gemini with the prompt (Appendix A.1.2) to extract sound-producing entities likely to appear in the visual scene. We sample one entity and ask: "*Is/Are the* $\{$*object*$\}$ *visible in the video? Answer Yes or No.*" Together with the intact video, this query is sent to Gemini for the reference label, ensuring the answer is grounded solely in visual evidence. ***For the audio-visual matching task***, we present the intact audio-visual pair to Gemini with the prompt: "*Are the audio and visual contents describing the same scene or event? Answer Yes or No.*" The obtained label indicates cross-modal

coherence, providing supervision for alignment. Unlike the prior two tasks, this one explicitly fuses both modalities to assess semantic alignment, thus regularizing joint representations.

Table 1 summarizes the data distribution across the three tasks. With 7,944 question-answering (QA) pairs in total, we maintain a reasonably balanced positive/negative ratio per task. Of note, all training samples are stored intact (original $A$, $V$, query $q$, and Gemini label $y$). Modality corruption (e.g., noising $\mathcal{M}_{\mathsf{A}}$ for $t = \mathsf{A}$) is applied dynamically during forward passes rather than at dataset creation. This decouples annotation from augmentation, enabling diverse corruptions across epochs without regenerating data.

### 4.2 Modality Intervention

We now introduce a modality intervention mechanism to generate heterogeneous input conditions for our method. Standard GRPO samples multiple responses from the same input, which is insufficient for modality-faithful behavior Shao et al. (2024). To address it, we systematically perturbs task-relevant and task-irrelevant modalities based on task type $t$, producing conditions that jointly enable optimizing correctness, robustness, and sensitivity within a unified framework.

First, to generate corrupted modality inputs, we adopt Eq. (1) and follow Chaubey et al. (2026) to fix $s = 500$. The corrupted audio and video inputs are thus defined as:

$$\tilde{A} = \sqrt{\bar{\alpha}_s}\, A + \sqrt{1 - \bar{\alpha}_s}\, \epsilon_A, \tilde{V} = \sqrt{\bar{\alpha}_s}\, V + \sqrt{1 - \bar{\alpha}_s}\, \epsilon_V, \tag{4}$$

where $\epsilon_A, \epsilon_V \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I})$. Building upon Eq. (4), we further expand each original training instance $(A, V, q, t, y)$ into a set of task-specific conditions, creating the heterogeneous rollouts required for group-wise optimization. Specifically, for single-modality tasks, i.e., $t = \mathsf{A}/\mathsf{V}$, we generate three conditions: the intact pair $(A, V)$ to maintain correctness, the pair with the task-irrelevant modality corrupted to augment robustness, and the pair with the task-relevant modality corrupted to enforce sensitivity (where the answer flips to "No"). For example, these correspond to $(A, V)$, $(A, \tilde{V})$, and $(\tilde{A}, V)$, respectively, for $t = \mathsf{A}$. For the audio-visual matching task ($t = \mathsf{AV}$), we retain only the intact input pair $(A, V)$. Ablation studies (see Sec. 5.3) show that introducing corrupted conditions for this task yields negligible performance gains; we therefore omit them for training efficiency.

### 4.3 Task-Aware Group Routing

We organize the heterogeneous conditions from modality intervention into behavioral groups to jointly optimize correctness, robustness, and sensitivity. The routing mechanism assigns each condition a task-dependent group and a group-specific reference answer, enabling per-group advantage normalization.

Concretely, the intact condition $(A, V)$ always forms the *Intact* group ($\mathrm{Int}$), preserving complete evidence for correctness. For single-modality tasks ($t = \mathsf{A}/\mathsf{V}$), the two corrupted variants are routed complementarily: the one with the *task-irrelevant* modality corrupted serves as the *Robustness* group ($\mathrm{Rob}$), and the one with the *task-relevant* modality corrupted constitutes the *Sensitivity* group ($\mathrm{Sen}$). For the audio-visual matching task ($t = \mathsf{AV}$), only $\mathrm{Int}$ is retained. This mapping takes the form:

$$\mathcal{R}(t) = \{(A, V) \to \mathrm{Int}\} \cup \begin{cases} \left\{ \begin{matrix} (A, \tilde{V}) \to \mathrm{Rob}, \\ (\tilde{A}, V) \to \mathrm{Sen} \end{matrix} \right\}, & t = \mathsf{A}, \\ \left\{ \begin{matrix} (\tilde{A}, V) \to \mathrm{Rob}, \\ (A, \tilde{V}) \to \mathrm{Sen} \end{matrix} \right\}, & t = \mathsf{V}, \\ \{(A, V) \to \mathrm{Int}\}, & t = \mathsf{AV}, \end{cases} \tag{5}$$

yielding the valid group sets $\mathcal{G}_{\mathsf{A}} = \mathcal{G}_{\mathsf{V}} = \{\mathrm{Int}, \mathrm{Rob}, \mathrm{Sen}\}$ and $\mathcal{G}_{\mathsf{AV}} = \{\mathrm{Int}\}$. For the training dataset $\mathcal{D} = \{(A_i, V_i, q_i, t_i, y_i)\}_{i=1}^{N}$, the input condition $c_i^g$ for sample $i$ and group $g \in \mathcal{G}_{t_i}$ is the modality pair assigned by $\mathcal{R}(t_i)$ together with $q_i$.

Furthermore, the reference answer for each group strictly follows evidence availability. For $\mathrm{Int}$ and $\mathrm{Rob}$, task-relevant evidence remains intact, so the target is the teacher answer $y_i$; for $\mathrm{Sen}$, the task-relevant modality is corrupted, making the original evidence invalid, that is,

$$\hat{y}_i^g = \begin{cases} y_i, & g \in \{\mathrm{Int}, \mathrm{Rob}\}, \\ \texttt{No}, & g = \mathrm{Sen}. \end{cases} \tag{6}$$

This produces unambiguous supervision: Int and Rob optimize correctness and robustness under preserved evidence, while Sen penalizes unsupported affirmative predictions when task-relevant information is absent.

### 4.4 Group-Weighted Objective

Given the grouped conditions $c_i^g$, we derive the optimization objective that jointly optimizes correctness, robustness, and sensitivity. To be specific, for each sample $i \in \{1, \ldots, N\}$ and group $g \in \mathcal{G}_{t_i}$, we sample $K_{t_i,g}$ responses from the old policy $\pi_{\theta_{\text{old}}}$: $o_{i,k}^g \sim \pi_{\theta_{\text{old}}}(\cdot \mid c_i^g)(k = 1, \ldots, K_{t_i,g})$, and evaluate them against $\hat{y}_i^g$ using a deterministic, regular-expression-based scoring rule, yielding binary rewards $r_{i,k}^g = \mathbb{I}[o_{i,k}^g = \hat{y}_i^g]$.

Then, the three groups (Int, Rob, Sen) represent distinct behavioral objectives (see Eq. (5)), so pooling their rewards would create harmful cross-objective competition. We therefore normalize advantages independently within each group. For every $i$ and $g$, we compute the mean $\mu_i^g$ and standard deviation $\sigma_i^g$ over the $K_{t_i,g}$ responses from condition $c_i^g$, and define the group-wise advantage as $\widehat{A}_{i,k}^g = \frac{r_{i,k}^g - \mu_i^g}{\sigma_i^g + \varepsilon}$. This ensures each rollout is compared only against responses sharing the same condition and objective, decoupling reward distributions across groups.

We next apply the clipped GRPO objective (see Eqs. (2) and (3)) with these advantages, the token-level loss for the $k$-th rollout in group $g$ is

$$\mathcal{L}_{i,k,m}^g(\theta) = -\min\left(\rho_{i,k,m}^g(\theta)\widehat{A}_{i,k}^g, \text{clip}\Big(\rho_{i,k,m}^g(\theta), 1-\epsilon, 1+\epsilon\Big)\widehat{A}_{i,k}^g\right), \tag{7}$$

where $\rho_{i,k,m}^g(\theta)$ is the policy probability ratio at token $m$. As such, the per-rollout loss is $\mathcal{L}_{i,k}^g(\theta) = \frac{1}{M_{i,k}^g}\sum_{m=1}^{M_{i,k}^g} \ell_{i,k,m}^g(\theta)$. Further, averaging over the $K_{t_i,g}$ rollouts yields the sample-level loss:

$$\mathcal{L}_{i,g}(\theta) = \frac{1}{K_{t_i,g}} \sum_{k=1}^{K_{t_i,g}} \ell_{i,k}^g(\theta) + \beta \mathbb{D}_{\text{KL}}\left(\pi_\theta \| \pi_{\text{ref}}\right). \tag{8}$$

Finally, the overall TAG-GRPO objective combines the losses per sample with task–group weights $w_{t_i,g}$ and averages over all $N$ samples:

$$\mathcal{L}_{\text{TAG-GRPO}}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \frac{\sum_{g \in \mathcal{G}_{t_i}} w_{t_i,g}\, \mathcal{L}_{i,g}(\theta)}{\sum_{g \in \mathcal{G}_{t_i}} w_{t_i,g}}. \tag{9}$$

These weights control the contribution of each behavioral objective, balancing correctness, robustness, and sensitivity without conflating their reward distributions.

## 5 Experiments

### 5.1 Experimental Settings

**Benchmarks.** We evaluate on AVHBench Sung-Bin et al. (2025) and CMM Leng et al. (2024a), which test complementary audio-visual hallucinations. All models receive separate audio and muted-video inputs. AVHBench covers audio-driven visual hallucination (ADVH), video-driven audio hallucination (VDAH), and audio-visual matching (AVM); we report accuracy, precision, recall, and F1. CMM assesses spurious cross-modal correlations (including vision–language (VL), audio–language (AL), and vision–audio–language (VAL)) and unimodal over-reliance (visual, audio, language), reporting perception accuracy (PA) and hallucination resistance (HR), averaged over the six subsets.

**Models and Baselines.** We apply all trainable methods to two omni-modal backbones: Qwen2.5-Omni-7B Xu et al. (2025a) and MiniCPM-o-2.6 Team (2025). The comparison includes inference with the base model (Prompting), DPO Rafailov et al. (2023), OmniDPO Chen et al. (2026), MoD-DPO Chaubey et al. (2026), and standard GRPO Shao et al. (2024). For context, we also report results

Table 2: Evaluation results on AVHBench Sung-Bin et al. (2025), comparing Qwen2.5-Omni-7B and MiniCPM-o-2.6 trained with different strategies against other models without additional training.

| Model/Method | Audio-driven Video Hallucination | | | | Video-driven Audio Hallucination | | | | AV Matching | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc (↑) | Pre (↑) | Rec (↑) | F1 (↑) | Acc (↑) | Pre (↑) | Rec (↑) | F1 (↑) | Acc (↑) | Pre (↑) | Rec (↑) | F1 (↑) |
| Nemotron-3-Nano-Omni | 73.77 | 76.63 | 68.25 | 72.20 | 68.25 | 73.88 | 56.38 | 63.96 | 65.67 | 72.90 | 49.89 | 59.24 |
| Qwen3-Omni-30B-A3B-Instruct | 84.15 | **83.54** | 85.01 | 84.27 | 84.15 | 79.52 | **91.96** | 85.29 | 82.68 | 79.00 | **89.02** | 83.71 |
| Gemini-3.1-pro-preview | **85.12** | 82.98 | **88.38** | **85.59** | **86.59** | **85.15** | 88.65 | **86.86** | **90.09** | **96.31** | 83.37 | **89.37** |
| **Qwen2.5-Omni-7B** | | | | | | | | | | | | |
| Prompting | 82.22 | 81.41 | 83.42 | 82.40 | 74.80 | 67.13 | 97.12 | 79.39 | 70.63 | 97.08 | 42.54 | 59.15 |
| + DPO | 84.60 | 87.40 | 80.78 | 83.96 | **81.18** | **74.24** | 95.45 | **83.52** | 71.22 | **97.84** | 43.39 | 60.12 |
| + OmniDPO | 83.01 | 84.00 | 81.48 | 82.72 | 76.81 | 69.14 | 96.77 | 80.66 | 70.58 | 97.54 | 42.22 | 58.93 |
| + MODDPO | 82.39 | 81.37 | **83.95** | 82.64 | 73.23 | 65.65 | **97.38** | 78.42 | 77.24 | 95.87 | 56.93 | 71.44 |
| + GRPO | 84.51 | 85.74 | 82.72 | 84.20 | 79.30 | 71.92 | 96.07 | 82.26 | 80.33 | 93.44 | 65.25 | 76.84 |
| **+ Ours** | **85.04** | **88.25** | 80.78 | **84.35** | 80.31 | 73.30 | 95.28 | 82.86 | **80.38** | 92.66 | **65.99** | **77.09** |
| Compared to Prompting | +2.82 | +6.84 | -2.64 | +1.95 | +5.51 | +6.17 | -1.84 | +3.47 | +9.75 | -4.42 | +23.45 | +17.94 |
| **MiniCPM-o-2.6** | | | | | | | | | | | | |
| Prompting | 81.07 | 80.24 | 82.36 | 81.29 | 78.34 | 76.09 | 82.60 | 79.21 | 68.55 | 85.95 | 44.35 | 58.51 |
| + DPO | 86.18 | 86.74 | 85.36 | 86.04 | 80.48 | 80.07 | 81.12 | 80.59 | 74.31 | 88.91 | 55.54 | 68.37 |
| + OmniDPO | 82.39 | 80.94 | 84.66 | 82.76 | 80.09 | 78.06 | 83.65 | 80.76 | 71.00 | **89.72** | 47.44 | 62.06 |
| + MODDPO | 83.36 | 82.25 | 85.01 | 83.61 | 81.09 | 80.26 | 82.43 | 81.33 | 73.24 | 88.93 | 53.09 | 66.49 |
| + GRPO | **86.27** | 84.89 | **88.18** | **86.51** | **81.79** | 77.48 | **89.60** | **83.10** | 68.18 | 61.80 | **95.20** | 74.95 |
| **+ Ours** | 85.56 | **91.38** | 78.48 | 84.44 | 79.69 | **80.84** | 77.80 | 79.29 | **76.81** | 79.01 | 73.03 | **75.90** |
| Compared to Prompting | +4.49 | +11.14 | -3.88 | +3.15 | +1.35 | +4.75 | -4.80 | +0.08 | +8.26 | -6.94 | +28.68 | +17.39 |

for Nemotron-3-Nano-Omni-30B-A3B-Reasoning-BF16 Deshmukh et al. (2026) (abbreviated as Nemotron-3-Nano-Omni), Qwen3-Omni-30B-A3B-Instruct Xu et al. (2025b), and Gemini-3.1-Pro-Preview Google (2026). Because these systems differ in model size and training data, their results are reference points rather than controlled training baselines.

**Implementation Details.** All trainable methods use LoRA Hu et al. (2022) with rank $r_{\text{LoRA}} = 8$ and scaling factor $\alpha_{\text{LoRA}} = 32$ on all linear layers via Swift Zhao et al. (2024). We use learning rates $5{\times}10^{-6}$ (Qwen2.5) and $5{\times}10^{-5}$ (MiniCPM). We train for 1 epoch using a batch size of 1, a warmup ratio of 0.05, and a cosine learning-rate schedule. Training is conducted on eight NVIDIA A800 GPUs with 40 GB of memory per GPU. The GRPO rollout budget is 8. For $t{\in}\{\mathsf{A},\mathsf{V}\}$, group loss weights are $(w_{\text{Int}}, w_{\text{Rob}}, w_{\text{Sen}}) = (0.7, 0.2, 0.1)$; for AV, only the intact condition is used, i.e., $(w_{\mathsf{AV},\text{Int}}, w_{\mathsf{AV},\text{Rob}}, w_{\mathsf{AV},\text{Sen}}) = (1.0, 0.0, 0.0)$, since corrupting either modality would invalidate the matching label.

### 5.2 Results Analysis

This section evaluates TAG-GRPO on hallucination-focused and general audio-visual benchmarks. All reported values are point estimates.

**Results on AVHBench.** Table 2 summarizes the AVHBench results. On Qwen2.5-Omni-7B, TAG-GRPO improves over Prompting across all subsets, with the largest gains in AVM (+9.75 Acc, +17.94 F1). Compared to GRPO, the differences are modest (+0.05–1.01 Acc) and lack statistical reliability. On MiniCPM-o-2.6, TAG-GRPO substantially boosts AVM (+8.63 Acc over GRPO) but falls behind GRPO on ADVH ($-0.71$ Acc) and VDAH ($-2.10$ Acc). Hence, TAG-GRPO strengthens audio-visual matching yet does not uniformly surpass GRPO on modality-specific hallucination tasks. Precision–recall analysis further reveals that TAG-GRPO raises precision while lowering recall on ADVH and VDAH for both backbones. On AVM, recall surges (e.g., +23.45 for Qwen2.5-Omni-7B) at the cost of precision. F1 improvements should therefore be interpreted together with these trade-offs.

**Results on CMM.** On CMM (Table 3), TAG-GRPO improves overall hallucination resistance (HR) over Prompting and GRPO for both backbones. For Qwen2.5-Omni-7B, HR rises from 71.42 (Prompting) / 73.92 (GRPO) to 76.00, though DPO achieves a slightly higher HR (76.83). On MiniCPM-o-2.6, HR jumps from 78.17 (Prompting) / 80.75 (GRPO) to 91.50, with gains in every subset. However, these HR improvements come at the cost of perception accuracy (PA): overall PA drops by 1.59 (Qwen2.5-Omni-7B) and 4.00 (MiniCPM-o-2.6) points. The current experiments do not isolate whether this trade-off arises from the group-specific targets, teacher annotations, judge rewards, or another training factor.

Table 3: Evaluation results on CMM Leng et al. (2024a), comparing Qwen2.5-Omni-7B and MiniCPM-o-2.6 trained with different strategies against other models without additional training.

| Model/Method | Spurious Inter-modality Correlation | | | | | | Uni-modality Over-reliance | | | | | | Overall | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | VL | | AL | | VAL | | Visual Dom. | | Audio Dom. | | Lang. Dom. | | | |
| | PA (↑) | HR (↑) | PA (↑) | HR (↑) | PA (↑) | HR (↑) | PA (↑) | HR (↑) | PA (↑) | HR (↑) | PA (↑) | HR (↑) | PA (↑) | HR (↑) |
| Nemotron-3-Nano-Omni | **94.5** | 87.5 | 92.0 | 52.0 | 96.0 | 76.5 | 89.0 | 52.0 | 92.5 | 69.0 | 89.5 | 70.0 | 92.25 | 67.83 |
| Qwen3-Omni-30B-A3B-Instruct | 93.0 | **90.5** | **93.0** | 91.0 | **97.5** | **90.5** | **95.0** | 69.5 | **96.5** | 55.5 | **92.5** | 71.0 | **94.58** | 78.00 |
| Gemini-3.1-pro-preview | 90.0 | 90.0 | 86.0 | **93.0** | 93.0 | 84.0 | 92.0 | **90.0** | 89.5 | **82.0** | 81.5 | **90.0** | 88.67 | **88.17** |
| **Qwen2.5-Omni-7B** | | | | | | | | | | | | | | |
| Prompting | 94.0 | 90.0 | **95.5** | 75.5 | 96.0 | 90.5 | 95.5 | 56.5 | **95.0** | 40.0 | 86.0 | 76.0 | 93.67 | 71.42 |
| + DPO | 93.0 | **94.0** | 92.5 | 79.0 | 92.0 | 95.0 | 94.0 | **65.0** | 94.5 | **48.5** | 84.5 | **79.5** | 91.75 | **76.83** |
| + OmniDPO | 93.5 | 91.0 | 95.5 | 77.5 | 95.0 | 93.0 | 95.0 | 58.0 | 95.0 | 41.0 | 85.5 | 77.5 | 93.25 | 73.00 |
| + MODDPO | **94.5** | 90.0 | 95.5 | 74.5 | **96.5** | 91.5 | **96.0** | 56.0 | 95.0 | 40.0 | **87.0** | 76.5 | **94.08** | 71.42 |
| + GRPO | 94.0 | 91.5 | 94.0 | 78.0 | 95.0 | 93.0 | 94.0 | 59.5 | 95.0 | 44.0 | 85.5 | 77.5 | 92.92 | 73.92 |
| **+ Ours** | 93.0 | 93.0 | 93.5 | **79.5** | 93.0 | **95.5** | 94.0 | 62.5 | 94.5 | 47.0 | 84.5 | 78.5 | 92.08 | 76.00 |
| Compared to Prompting | -1.0 | +3.0 | -2.0 | +4.0 | -3.0 | +5.0 | -1.5 | +6.0 | -0.5 | +7.0 | -1.5 | +2.5 | -1.59 | +4.58 |
| **MiniCPM-o-2.6** | | | | | | | | | | | | | | |
| Prompting | 88.0 | 95.0 | **93.5** | 74.5 | 93.0 | 91.0 | 90.0 | 75.0 | 89.0 | 53.0 | 81.0 | 80.5 | **89.08** | 78.17 |
| + DPO | 88.0 | 96.5 | 89.0 | 83.5 | 91.0 | 92.0 | 88.5 | 82.5 | 87.5 | 63.0 | 80.0 | 81.5 | 87.33 | 83.17 |
| + OmniDPO | 87.5 | 95.0 | 91.0 | 77.5 | **94.0** | 89.5 | 91.5 | 78.0 | 87.0 | 55.0 | 80.5 | 81.0 | 88.58 | 79.33 |
| + MODDPO | 87.5 | 95.5 | 90.0 | 81.0 | 93.0 | 89.0 | **91.5** | 80.5 | 87.0 | 54.0 | 79.0 | 81.5 | 88.00 | 80.25 |
| + GRPO | **91.5** | 96.5 | 88.5 | 82.5 | 92.0 | 94.5 | 88.0 | 62.0 | **89.5** | 73.0 | **82.5** | 76.0 | 88.67 | 80.75 |
| **+ Ours** | 87.0 | **98.0** | 91.0 | **89.0** | 88.0 | **98.0** | 83.0 | **89.5** | 83.5 | **90.0** | 78.0 | **84.5** | 85.08 | **91.50** |
| Compared to Prompting | -1.0 | +3.0 | -2.5 | +14.5 | -5.0 | +7.0 | -7.0 | +14.5 | -5.5 | +37.0 | -3.0 | +4.0 | -4.00 | +13.33 |

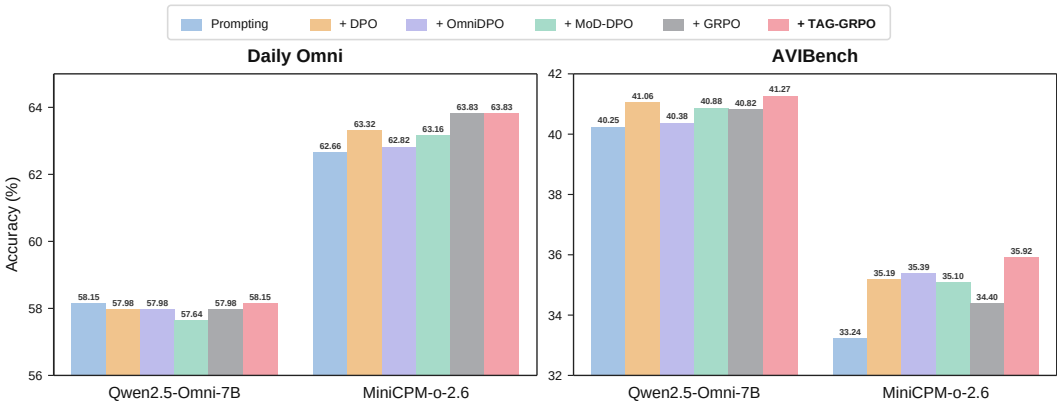



Figure 3: Accuracy comparison of TAG-GRPO and baseline methods on Daily Omni Zhou et al. (2025) and AVIBench Wang et al. (2026) using Qwen2.5-Omni-7B and MiniCPM-o-2.6. Higher is better.

**Non-hallucination Benchmarks.** Figure 3 reports average accuracy on Daily Omni Zhou et al. (2025) and AVIBench Wang et al. (2026). On Daily Omni, Qwen2.5-Omni-7B remains at $58.15\%$, while MiniCPM-o-2.6 increases from $62.66\%$ to $63.83\%$, a difference of $1.17$ points. On AVIBench, Qwen2.5-Omni-7B increases from $40.25\%$ to $41.27\%$, and MiniCPM-o-2.6 increases from $33.24\%$ to $35.92\%$, corresponding to differences of $1.02$ and $2.68$ points. Relative to GRPO, the AVIBench differences are $+0.45$ points for Qwen2.5-Omni-7B and $+1.52$ points for MiniCPM-o-2.6. These two benchmarks show no observed decrease in average accuracy for the evaluated settings, but they do not establish preservation of capabilities outside the tested tasks.

### 5.3 Ablation and Sensitivity Analyses

**Ablation Study.** The ablations use Qwen2.5-Omni-7B and report point estimates on AVHBench and CMM. Table 4 varies the training task types, active behavior groups, and corrupted modalities. Because these are distinct factors, comparisons are organized by ablation block rather than pooled across all rows. **Effect of Training Task Types.** The first block of Table 4 separates modality-specific and relation-level training. Training only on A and V yields ADVH, VDAH, and HR values of $85.21$, $80.83$, and $78.25$, but an AVM value of $67.06$. Training only on AV raises AVM to $83.16$ and PA to $95.92$, while ADVH, VDAH, and HR decrease to $75.62$, $66.07$, and $63.50$. The full configuration lies between these task-specific settings: it recovers ADVH, VDAH, and HR relative to AV-only training and recovers AVM relative to $\{\mathsf{A}, \mathsf{V}\}$-only training. This observation indicates that modality-specific

Table 4: Ablation results of Qwen2.5-Omni-7B under different task, behavior group, and modality corruption configurations. Here, I., R., and S. denote Intact, Robustness, and Sensitivity groups, while $(\tilde{A}, V)$ and $(A, \tilde{V})$ denote audio-corrupted and video-corrupted conditions.

| Tasks type | Group | Corruption | AVHBench | | | CMM | |
|---|---|---|---|---|---|---|---|
| | | | ADVH | VDAH | AVM | PA | HR |
| A, V | I. + R. + S. | $(\tilde{A}, V), (A, \tilde{V})$ | 85.21 | 80.83 | 67.06 | 91.0 | 78.25 |
| AV | I. + R. + S. | $(\tilde{A}, V), (A, \tilde{V})$ | 75.62 | 66.07 | 83.16 | 95.92 | 63.50 |
| A, V, AV | I. + R. | $(\tilde{A}, V), (A, \tilde{V})$ | 84.86 | 79.61 | 76.76 | 92.58 | 75.00 |
| A, V, AV | I. + S. | $(\tilde{A}, V), (A, \tilde{V})$ | 84.77 | 80.87 | 78.04 | 91.83 | 76.50 |
| A, V, AV | I. + R. + S. | $(\tilde{A}, V)$ | 85.04 | 80.70 | 78.68 | 92.08 | 76.75 |
| A, V, AV | I. + R. + S. | $(A, \tilde{V})$ | 85.21 | 80.66 | 77.35 | 91.83 | 77.08 |
| A, V, AV | I. + R. + S. | $(\tilde{A}, V), (A, \tilde{V})$ | 85.04 | 80.31 | 80.38 | 92.08 | 76.00 |

and matching tasks contribute to different evaluation dimensions; none of the task subsets is uniformly best. **Effect of Behavior Groups.** The second block compares the full configuration with variants that omit one behavior group for $t \in \{\mathsf{A}, \mathsf{V}\}$. Removing the Sensitivity group ($\mathrm{Int} + \mathrm{Rob}$) changes ADVH, VDAH, AVM, PA, and HR by $-0.18$, $-0.70$, $-3.62$, $+0.50$, and $-1.00$ points relative to the full configuration. Removing the Robustness group ($\mathrm{Int} + \mathrm{Sen}$) changes the same metrics by $-0.27$, $+0.56$, $-2.34$, $-0.25$, and $+0.50$ points. Both variants have lower AVM, whereas their effects on VDAH, PA, and HR are mixed. This ablation is consistent with retaining both groups for the reported AVM result, but it does not show that either group improves every metric. **Effect of Corruption Coverage.** The third block varies whether training includes corrupted audio, corrupted video, or both for the modality-specific tasks. The two-corruption configuration yields AVM of $80.38$, compared with $78.68$ for audio corruption alone and $77.35$ for video corruption alone. Conversely, each single-corruption variant is higher on VDAH and HR, and the video-corruption-only variant is $0.17$ points higher on ADVH. Thus, the two-corruption configuration is associated with a higher AVM value in this ablation, but does not dominate the single-corruption variants on all metrics.

**Sensitivity to Audio–Visual Rollout Allocation.** Table 5 varies the allocation of eight AV rollouts among intact, audio-corrupted, and video-corrupted conditions. Assigning all eight rollouts to $\mathrm{Int}$ gives the highest ADVH, VDAH, and AVM values among the three settings. Relative to the $6/1/1$ allocation, the differences are $+0.27$, $+0.18$, and $+2.87$ points, respectively. The CMM differences are mixed: the all-intact allocation has PA $0.09$ points lower and HR $0.08$ points higher than $6/1/1$, while its HR is $0.67$ points lower than $4/2/2$. The all-intact setting is therefore selected on the basis of its AVHBench results, especially AVM; it is not uniformly best on CMM. Corrupting either stream changes the relation evaluated by an audio–visual matching question. The allocation sensitivity is consistent with avoiding a fixed negative target for such corrupted pairs. However, because the table covers only three allocations and one backbone, this observation is specific to the evaluated configuration.

Table 5: Sensitivity to rollout allocation for AV training with Qwen2.5-Omni-7B. $(A, V)$, $(\tilde{A}, V)$ and $(A, \tilde{V})$ denote intact, audio-corrupted, and video-corrupted conditions.

| Audio–Visual Task Rollout Allocation | | | AVHBench | | | CMM | |
|---|---|---|---|---|---|---|---|
| $(A, V)$ | $(\tilde{A}, V)$ | $(A, \tilde{V})$ | ADVH | VDAH | AVM | PA | HR |
| 4 | 2 | 2 | 84.51 | 80.22 | 74.04 | 91.42 | 76.67 |
| 6 | 1 | 1 | 84.77 | 80.13 | 77.51 | 92.17 | 75.92 |
| 8 | 0 | 0 | 85.04 | 80.31 | 80.38 | 92.08 | 76.00 |

## 6 Conclusion

In this paper, we proposed TAG-GRPO, a task-aware grouped policy optimization framework that extends GRPO with modality-conditioned counterfactual rollouts to explicitly encourage modality-faithful reasoning. By organizing responses into Intact, Robustness, and Sensitivity groups and applying group-wise advantage normalization, TAG-GRPO jointly optimizes correctness, robustness to irrelevant modality perturbations, and sensitivity to the disruption of task-relevant evidence. Extensive experiments on AVHBench and CMM show that TAG-GRPO effectively reduces cross-modal hallucinations and strengthens audio-visual grounding, highlighting the critical importance of optimizing modality-faithful behaviors beyond answer correctness alone.

## AI USE STATEMENT

We used GPT to assist with translation and English-language polishing of the manuscript. The authors manually reviewed the AI-assisted text for accuracy and consistency with the intended meaning.

For training-data construction, we used Gemini-3.1-Pro-Preview to extract candidate objects or sound sources and generate reference answers for training examples derived from VGGSound. The annotation procedure is described in Section 4.1, and the object-mining prompts are provided in Appendix A.1. This automated annotation workflow was adopted to reduce manual labeling costs; the generated candidates and reference answers were not manually corrected.

Generative AI was not used for method design, analysis or interpretation of experimental results, or figure creation. The authors take full responsibility for the final manuscript and the research, including the use of model-generated training annotations.

## REFERENCES

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, et al. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*, 2025.

Rui Cai, Bangzheng Li, Xiaofei Wen, Muhao Chen, and Zhe Zhao. Diagnosing and mitigating modality interference in multimodal large language models. *arXiv preprint arXiv:2505.19616*, 2025.

Ashutosh Chaubey, Jiacheng Pang, and Mohammad Soleymani. Mod-dpo: Towards mitigating cross-modal hallucinations in omni llms using modality decoupled preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18284–18294, 2026.

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

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Chao Sun, Rongzhou Zhang, Guanyu Zhou, and Lijie Wen. Omnidpo: A preference optimization framework to address omni-modal hallucination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20172–20180, 2026.

Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Yaoting Wang, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Avtrustbench: Assessing and enhancing reliability and robustness in audio-visual llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1590–1601, 2025.

Sangyun Chung, Se Yeon Kim, Youngchae Chee, and Yong Man Ro. Mad: Modality-adaptive decoding for mitigating cross-modal hallucinations in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40175–40185, 2026.

Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, Matthieu Le, Tyler Poon, Danial Mohseni Taheri, Ilia Karmanov, Guilin Liu, Jarno Seppanen, Arushi Goel, et al. Nemotron 3 nano omni: Efficient and open multimodal intelligence. *arXiv preprint arXiv:2604.24954*, 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Google. Gemini 3.1 Pro Preview. https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview, 2026. Model ID: `gemini-3.1-pro-preview`; accessed June 17, 2026.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14375–14385, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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*, 2024a. URL https://arxiv.org/abs/2410.12787.

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.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023a.

Xuanchen Li, Yuheng Lu, Chenrui Cui, Tianrui Wang, Zikang Huang, Yu Jiang, Long Zhou, Longbiao Wang, and Jianwu Dang. Separate first, fuse later: Mitigating cross-modal interference in audio-visual llms reasoning with modality-specific chain-of-thought. *arXiv preprint arXiv:2605.09906*, 2026.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023b.

Chenxi Liu, Tianyi Xiong, Yanshuo Chen, Ruibo Chen, Yihan Wu, Junfeng Guo, Tianyi Zhou, and Heng Huang. Modality-balancing preference optimization of large multimodal models by adversarial negative mining. *arXiv preprint arXiv:2506.08022*, 2025a.

Shujun Liu, Siyuan Wang, Zejun Li, Jianxiang Wang, Cheng Zeng, and Zhongyu Wei. Ovip: Online vision-language preference learning. *arXiv e-prints*, pp. arXiv–2505, 2025b.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

MTEB. VGGSound. https://modelscope.cn/datasets/MTEB/VGGSound, 2026. Hosted on ModelScope; accessed June 17, 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

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. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018.

Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass. Omni-r1: Do you really need audio to fine-tune your audio llm? *arXiv preprint arXiv:2505.09439*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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

Kim Sung-Bin, Oh Hyun-Bin, Lee Jung-Mok, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. In *International Conference on Learning Representations*, volume 2025, pp. 24244–24271, 2025.

OpenBMB MiniCPM-o Team. Minicpm-o 2.6: A gpt-4o level mllm for vision, speech, and multimodal live streaming on your phone. *Accessed: January*, 31:2025, 2025.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8078–8088, 2024.

Yaoting Wang, Ziyi Zhang, Wenming Tu, Shaoxuan Xu, Wenjie Du, Cheng Liang, Weijun Wang, Yuanchao Li, Guangyao Li, Hao Fei, Yuanchun Li, Henghui Ding, and Yunxin Liu. Avi-bench: Toward human-like audio-visual intelligence of omni-mllms. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

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

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10610–10620, 2025.

Zhangyuan Yu, Wanran Sun, Guangjing Yang, Xiaohu Wu, and Qicheng Lao. Cfpo: Counterfactual policy optimization for multimodal reasoning, 2026. URL https://arxiv.org/abs/2606.23206.

Chenshuang Zhang, Kyeong Seon Kim, Chengxin Liu, and Tae-Hyun Oh. Svhalluc: Benchmarking speech-vision hallucination in audio-visual large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25304–25314, 2026.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pp. 543–553, 2023.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift:a scalable lightweight infrastructure for fine-tuning, 2024. URL https://arxiv.org/abs/2408.05517.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023.

Xu Zheng, Chenfei Liao, Yuqian Fu, Kaiyu Lei, Yuanhuiyi Lyu, Lutao Jiang, Bin Ren, Jialei Chen, Jiawen Wang, Chengxin Li, et al. Mllms are deeply affected by modality bias. *arXiv preprint arXiv:2505.18657*, 2025.

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

# A PROMPT POOL

## A.1 OBJECT MINING TEMPLATES

**Prompt A.1.1: Visible Object Mining**

```
List visible non-human objects, entities, or events in the video that
    are likely capable of producing distinct, directly audible sounds
    and can be used as sound-source query labels.

Use only video as evidence. Prefer vehicles, animals, instruments,
    machines, alarms, phones, bells, doors, tools, running water, fire,
     or active events with an obvious sound source. Avoid silent or
    mostly static items such as furniture, clothes, plants, buildings,
    signs, screens, lights, sky, clouds, roads, walls, floors, scenery,
     text, logos, and generic background. Avoid people, human body
    parts, speech, voices, singing, laughter, and applause.

Return only a lowercase JSON array of short noun phrases.
```

**Object Mining Prompt-1.** This prompt extracts visible non-human objects, entities, or events that can plausibly produce distinctive sounds and serve as candidate labels for subsequent audibility questions. It uses only visual evidence and returns a lowercase JSON array of short noun phrases while excluding human-related content and visually irrelevant background elements.

**Prompt A.1.2: Audible Object Mining**

```
List concrete non-human physical sound sources that are directly
    audible in the audio and could plausibly be visible as objects or
    events in a video.

Use only audio as evidence. Return target labels such as vehicles,
    animals, instruments, machines, alarms, phones, bells, doors, tools
    , water sources, or weather events only when their sound is
    directly audible. Avoid people, human body parts, speech, voices,
    singing, laughter, applause, generic music, ambience, noise, wind,
    footsteps, impacts, and abstract sounds.

Return only a lowercase JSON array of short noun phrases.
```

**Object Mining Prompt-2.** This prompt extracts concrete non-human sound sources that are directly supported by the audio and can serve as candidate labels for subsequent visibility questions. It uses only acoustic evidence and returns a lowercase JSON array of short noun phrases while excluding human, speech-related, ambient, and abstract sounds.