HARPS: MITIGATING MULTIMODAL HALLUCINATIONS VIA HEAD-SELECTIVE AND ADAPTIVE PERCEPTUAL SINK REWEIGHTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models increasingly reason over multiple perceptual modalities, yet hallucination remains a persistent challenge. Recent work has shown that attention sink tokens can concentrate high-level perceptual and cross-modal information, motivating their use as inference-time intervention targets. We find, however, that these sink tokens behave differently across attention heads: unfaithful heads allocate more perceptual attention to sink positions, while their sink contributions are less semantically aligned with the input than those of faithful heads. Motivated by this finding, we propose **H**ead-Selective **A**daptive **R**eweighting for **P**erceptual **S**inks (HARPS), a training-free method that selectively reweights attention to perceptual sink tokens in faithful heads and adapts the layer-wise intervention to each input-prompt pair. Across four multimodal LLMs spanning different families, architectures, and scales, HARPS consistently improves both hallucination-focused and general multimodal reasoning benchmarks, while remaining substantially more efficient than other baseline training-free methods.

1 INTRODUCTION

Multimodal large language models (MLLMs) have made substantial progress in vision-language (Liu et al., 2023; Bai et al., 2025; Wang et al., 2024a; Agarwal et al., 2025; Clark et al., 2026) and audio-language understanding (Kong et al., 2024; Ghosh et al., 2026; Ding et al., 2025; Nam et al., 2026). More recently, unified models that jointly process multiple perceptual modalities have emerged (Xu et al., 2025a;b; Abdin et al., 2024; Comanici et al., 2025; Cheng et al., 2024; Zhan et al., 2024). These models must integrate evidence across modalities while grounding their responses in the perceptual information relevant to a given text prompt. However, hallucination remains a significant challenge in this setting, with models generating responses that are not faithfully grounded in the available perceptual evidence (Yang et al., 2025b; Nguyen et al., 2026; Selvakumar et al., 2026; Leng et al., 2026; Sung-Bin et al., 2025; Ye et al., 2025).

To mitigate hallucination without costly model retraining, prior works have explored contrastive decoding (Pang et al., 2025; Jung et al., 2026a; Chung et al., 2026; Wang et al., 2024b; Leng et al., 2024) and direct attention interventions (Huang et al., 2024; Yoo et al., 2026; Li et al., 2023b; Yang et al., 2025a). Recent studies further show that attention sink tokens can capture high-level perceptual and cross-modal information, motivating their use as intervention targets (Xiao et al., 2024; Sun et al., 2024; Luo et al., 2026; Jung et al., 2026b). However, existing sink-based methods typically treat these tokens uniformly across attention heads, leaving unclear whether their roles differ across heads.

This motivates us to examine whether perceptual sink tokens play the same role across different attention heads. We first estimate the faithfulness of each head using intervention-based attribution across multiple proxy corpora spanning audio and visual content, contrasting its contribution to hallucinated and grounded generation (Wang et al., 2023; Yang et al., 2025a). This separates the model into *unfaithful* heads, which are more strongly associated with hallucinated generation, and the remaining *faithful* heads. We then compare how perceptual sink tokens behave across the two groups and find a striking contrast: unfaithful heads allocate a larger fraction of their perceptual attention

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.

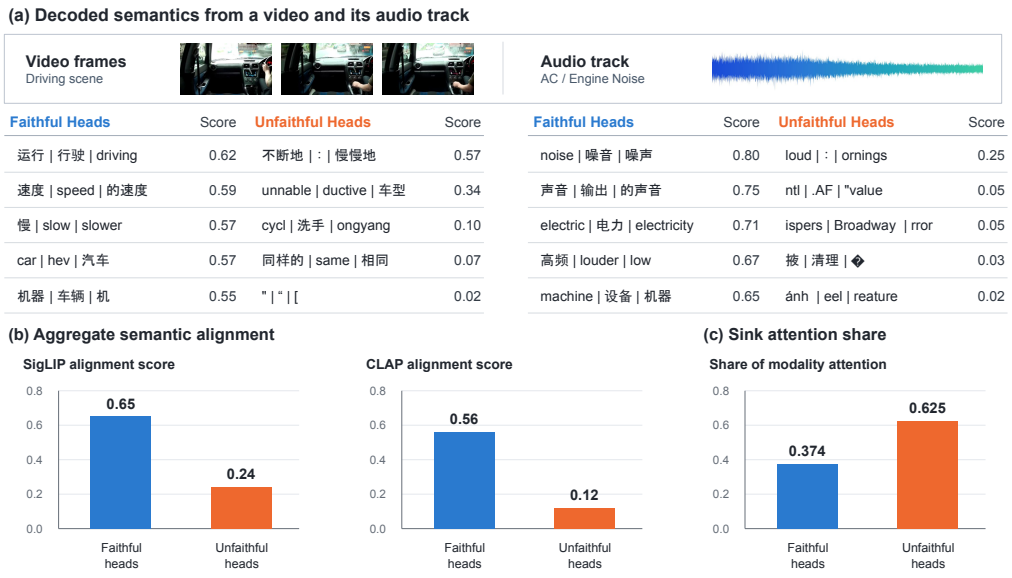


Figure 1: **Perceptual sink tokens behave differently across faithful and unfaithful attention heads.** **(a)** We decode sink-token contributions through individual heads using the OV-circuit view (Elhage et al., 2021) on Qwen2.5-Omni-7B; faithful heads recover more input-relevant concepts, including 行驶 (driving), 速度 (speed), 汽车 (car), 噪音 (noise), and 声音 (sound), with higher SigLIP (Zhai et al., 2023) and CLAP (Wu et al., 2023) alignment scores in the shown video and audio examples. **(b)** Aggregated over 300 VGGSounder examples, faithful heads achieve higher semantic alignment for both visual and audio inputs. **(c)** In contrast, unfaithful heads allocate a larger fraction of their perceptual attention to sink positions.

to sink positions, yet their head-level contributions at these positions are less semantically aligned with the perceptual input than those of faithful heads, as illustrated in Fig. 1. These findings suggest that interventions on perceptual sink tokens should account for the faithfulness of individual attention heads. Motivated by this observation, we propose **H**ead-Selective **A**daptive **R**eweighting for **P**erceptual **S**inks (HARPS), a training-free method for multimodal hallucination mitigation. Rather than reweighting attention to perceptual sink tokens uniformly across the model, HARPS uses the attribution-derived head partition to selectively reweight their attention in faithful heads, while leaving unfaithful heads unchanged.

We further observe that perceptual sink tokens exhibit distinct layer-wise distributions across inputs. In particular, sinks associated with audio, visual, and joint audio-visual information are distributed differently across decoder layers, and these distributions vary across examples. Moreover, the perceptual information relevant to generation depends on the text prompt: some prompts rely primarily on audio evidence, others on visual evidence, while some require information from both modalities. Based on these observations, we propose an adaptive reweighting strategy that adjusts the layer-wise intervention for each input-prompt pair. The layer-wise distributions of audio-, visual-, and audio-visual sinks determine three reweighting profiles, which are then combined according to the modality requirements of the text prompt into a single layer-wise schedule. This schedule is computed once from an initial unmodified prefill pass and then kept fixed throughout autoregressive decoding, adding little inference overhead; overall, HARPS remains substantially more efficient than the compared training-free methods.

We evaluate HARPS on four multimodal LLMs spanning two model families, dense and MoE architectures, and scales from 3B to 30B parameters. HARPS consistently improves hallucination-focused benchmarks as well as general reasoning across audio-video, image-audio, audio-only, and visual-only settings. Across the 24 accuracy-based model–benchmark combinations, 22 improve over the corresponding base model by at least 0.8 percentage points, with gains of up to 6.8 percentage points. On open-ended VGGSounder captioning, HARPS further reduces both CHAIR_S and CHAIR_I for all four models, achieving the best result in six of eight settings.

2 Related Work

Multimodal Large Language Models. Multimodal LLMs have progressed from vision-language models such as BLIP-2 (Li et al., 2023a), InstructBLIP (Dai et al., 2023), and LLaVA (Liu et al., 2023) to models that support increasingly diverse perceptual inputs. Video-oriented models such as Video-LLaMA (Zhang et al., 2023) and VideoLLaMA2 (Cheng et al., 2024) extend multimodal reasoning to temporal visual content, while more recent models including Qwen2.5-Omni (Xu et al., 2025a), Qwen3-Omni (Xu et al., 2025b), Phi-4 Multimodal (Abdin et al., 2024), VITA (Fu et al., 2026b), and OneLLM (Han et al., 2024) jointly process multiple perceptual modalities within a unified model. In this work, we propose HARPS, a training-free inference-time method for mitigating hallucination in multimodal LLMs that jointly process multiple perceptual modalities.

Mechanistic Analysis of Multimodal LLMs. Mechanistic analyses of multimodal LLMs have examined both the contributions of individual model components and the structure of internal attention. Intervention-based attribution methods localize model behavior to specific heads and layers (Wang et al., 2023; Meng et al., 2022), with AD-HH (Yang et al., 2025a) further identifying attention heads associated with hallucinated generation. A complementary line of work studies *attention sinks*, token positions that attract disproportionately large attention (Xiao et al., 2024; Sun et al., 2024; Zhang et al., 2025). In multimodal models, visual sink tokens have been linked to massive activations and high-level semantic information (Kang et al., 2025; Luo et al., 2026), while recent work on audio-visual models finds that sink tokens can capture modality-specific and shared information (Jung et al., 2026b). Our work connects these two lines by examining how perceptual sink tokens behave across attention heads with different levels of faithfulness.

Training-Free Multimodal Hallucination Mitigation. Training-free methods mitigate hallucination at inference time without modifying model parameters. One line of work uses contrastive decoding, where predictions from the original input are contrasted with those obtained from perturbed inputs or intermediate representations (Chuang et al., 2024; Leng et al., 2024; Wang et al., 2024b). This paradigm has been extended to video and audio-visual reasoning through methods such as AVCD (Jung et al., 2026a), MAD (Chung et al., 2026), and SEASON (Wu et al., 2026). Another line directly intervenes on attention during inference. AD-HH (Yang et al., 2025a) suppresses attention heads associated with hallucination, while recent methods reweight attention to perceptual sink tokens to improve grounding (Luo et al., 2026; Jung et al., 2026b). Our method builds on this direction by combining head-level faithfulness with adaptive reweighting of perceptual sink tokens.

3 Method

We propose Head-Selective Adaptive Reweighting for Perceptual Sinks (HARPS), a training-free method for mitigating hallucination in multimodal LLMs at inference time. HARPS first identifies attention heads according to their association with hallucinated generation (Sec. 3.1), and then selectively reweights attention to perceptual sink tokens in the remaining faithful heads (Sec. 3.2). The reweighting is further adapted across decoder layers for each input-prompt pair (Sec. 3.3). An overview of HARPS is shown in Fig. 2.

3.1 Identifying Faithful and Unfaithful Attention Heads

We first identify attention heads that contribute disproportionately to hallucinated rather than grounded generation. Following prior work on intervention-based attribution of internal model components (Wang et al., 2023; Meng et al., 2022; Michel et al., 2019), we ablate one attention head at a time and measure the resulting change in the model probability assigned to generated tokens. We then adopt the contrastive influence criterion of Yang et al. (2025a), which distinguishes heads that contribute more strongly to hallucinated generation than to grounded generation. For an attribution corpus p, let $\Delta_h^{(p)}[T]$ denote the average decrease in probability assigned to a set of token positions T after ablating head h. We define

$$s_h^{(p)} = \Delta_h^{(p)}[T_{\mathrm{hal}}] - \Delta_h^{(p)}[T_{\mathrm{gnd}}], \tag{1}$$

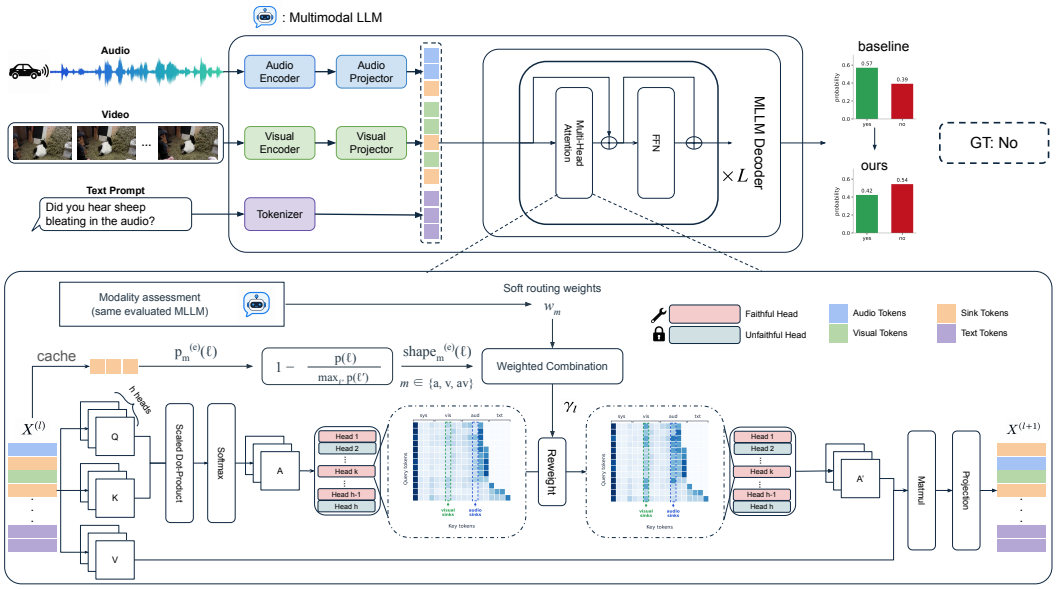


Figure 2: **Overview of HARPS.** HARPS uses an attribution-derived partition of attention heads into faithful and unfaithful groups. Given multimodal inputs and a text prompt, perceptual sink tokens are identified and selectively reweighted only in faithful heads. The reweighting strength is adapted across decoder layers using the example-specific sink-layer distributions together with routing weights obtained from prompt modality assessment. The resulting layer-wise schedule is applied during generation while unfaithful heads remain unchanged.

where T_{hal} and T_{gnd} denote token positions corresponding to hallucinated and grounded content, respectively. A positive $s_h^{(p)}$ indicates that head h contributes more strongly to hallucinated than grounded generation.

To obtain reliable grounded and hallucinated token positions for attribution, we construct examples for which these labels can be determined directly from corpus annotations. For each corpus $\mathcal{D}_p$ with label vocabulary $\mathcal{L}_p$, we prompt the model to generate a comma-separated list of labels drawn only from $\mathcal{L}_p$. Given the ground-truth label set $G \subseteq \mathcal{L}_p$, generated labels in G are treated as grounded and those outside G as hallucinated, yielding the corresponding token sets T_{gnd} and T_{hal}. This constrained generation avoids the additional parsing, matching heuristics, or judge models required to partition free-form responses. The exact prompting format and additional implementation details are provided in Appendix A.1.

We perform attribution over four proxy corpora spanning audio and visual content: ActivityNet (Caba Heilbron et al., 2015), YouTube-VOS (Xu et al., 2018), LibriSpeech (Panayotov et al., 2015), and AudioSet (Gemmeke et al., 2017). We classify a head as *unfaithful* if it has a positive hallucination-association score, $s_h^{(p)} > 0$, on at least one corpus, and as *faithful* otherwise. We denote the resulting sets by $\mathcal{H}_{\text{unfaithful}}$ and $\mathcal{H}_{\text{faithful}}$, respectively. Per-corpus attribution scores are provided in Appendix A.2.

3.2 Head-Selective Sink Reweighting

Recent studies have shown that attention sink tokens, beyond attracting disproportionately large attention (Xiao et al., 2024; Sun et al., 2024), can concentrate high-level perceptual and cross-modal information in multimodal LLMs (Kang et al., 2025; Luo et al., 2026; Jung et al., 2026b). Following Kang et al. (2025), we identify sink positions from dimensions with unusually large normalized activations. We extract the sink dimensions from the language backbone to avoid including dimensions specific to individual perceptual modalities; model-specific detection thresholds and additional details are provided in Appendix B. We denote the resulting set of sink positions by $\mathcal{S}$.

We find, however, that the semantic role of perceptual sink tokens differs substantially across faithful and unfaithful heads. To examine the contribution of individual heads at sink positions, we decode

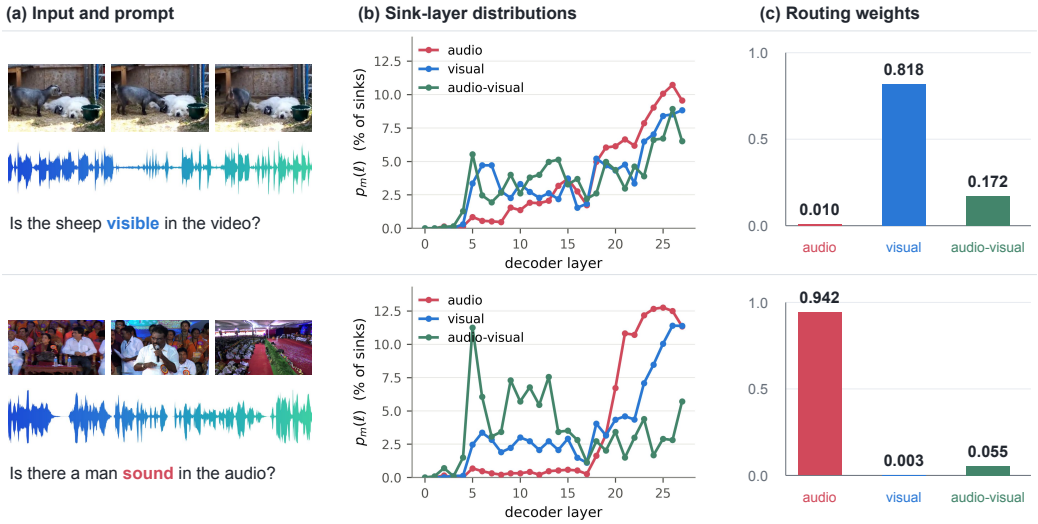


Figure 3: **Example-specific adaptive reweighting. (a)** Two representative audio-video inputs with text prompts requiring primarily visual evidence (top) and audio evidence (bottom). **(b)** The corresponding layer-wise distributions of audio-, visual-, and audio-visual sink tokens, which vary substantially across examples. **(c)** Routing weights produced by the same multimodal LLM from the text prompt, assigning most of the weight to the visual profile for the top example and to the audio profile for the bottom example.

each head's value contribution into the model vocabulary following the OV-circuit view (Elhage et al., 2021), and measure the semantic alignment of the decoded concepts with the perceptual input using SigLIP (Zhai et al., 2023) for visual inputs and CLAP (Wu et al., 2023) for audio inputs. As shown in Fig. 1, faithful heads yield concepts that are more semantically aligned with the perceptual input, whereas unfaithful heads more often produce fragmented or unrelated content. Interestingly, this difference is not explained by weaker attention to sink positions in unfaithful heads. On the contrary, unfaithful heads allocate a larger fraction of their perceptual attention to sink positions than faithful heads, averaging 0.625 versus 0.374.

These observations motivate a head-selective intervention. Rather than reweighting attention to perceptual sink tokens across all heads, HARPS increases attention to these tokens only in faithful heads, while leaving unfaithful heads unchanged. Specifically, let $A_{t,j}^{(\ell,h)}$ denote the attention weight from query position t to token position j in head h at layer ℓ. For a faithful head $h \in \mathcal{H}_{\text{faithful}}$ and a sink position $j \in \mathcal{S}$, we scale the corresponding attention weight by a factor $1+\gamma$:

$$\widetilde{A}_{t,j}^{(\ell,h)} = \begin{cases} (1+\gamma)A_{t,j}^{(\ell,h)}, & j \in \mathcal{S},\ h \in \mathcal{H}_{\text{faithful}}, \\ A_{t,j}^{(\ell,h)}, & \text{otherwise.} \end{cases} \tag{2}$$

The modified attention weights are then renormalized along each row to sum to one. Thus, the intervention selectively increases attention to perceptual sink tokens in faithful heads while leaving unfaithful heads unchanged.

3.3 Adaptive Reweighting

While a constant reweighting strength γ in Eq. 2 already provides a strong baseline (Table 4), we observe that the layer-wise distribution of perceptual sink tokens varies across examples and also differs according to their association with audio and visual information. The perceptual information relevant to generation further depends on the text prompt, with different prompts relying primarily on audio, visual, or joint audio-visual evidence. Fig. 3 shows representative examples in which both the layer-wise sink distributions and the prompt-dependent routing weights vary across input-prompt pairs. We therefore adapt the reweighting across layers for each input-prompt pair.

For each test example e, we first perform an unmodified prefill pass and use the resulting attention statistics to group perceptual sink tokens according to their association with audio and visual

information. Following Jung et al. (2026b), for each sink position j, we compute the modality dominance score $\mathrm{MDS}(j) = (\alpha_a(j) - \alpha_v(j))/(\alpha_a(j) + \alpha_v(j))$, where $\alpha_a(j)$ and $\alpha_v(j)$ denote the total attention received from audio and visual token positions, respectively. We classify a sink as audio-dominant when $\mathrm{MDS}(j) \geq 0.3$, visual-dominant when $\mathrm{MDS}(j) \leq -0.3$, and audio-visual otherwise, yielding the example-specific sets $\mathcal{S}_a^{(e)}$, $\mathcal{S}_v^{(e)}$, and $\mathcal{S}_{av}^{(e)}$.

For each sink type $m \in \{a, v, av\}$, let $p_m^{(e)}(\ell)$ denote the fraction of type-m sink positions that occur at decoder layer ℓ, with $\sum_\ell p_m^{(e)}(\ell) = 1$. We use this layer-wise distribution to define a reweighting profile

$$\mathrm{shape}_m^{(e)}(\ell) = 1 - \frac{p_m^{(e)}(\ell)}{\max_{\ell'} p_m^{(e)}(\ell')}. \tag{3}$$

For each sink type, layers containing relatively fewer sink positions receive stronger reweighting, while layers where that sink type is more concentrated receive less amplification. Since $p_m^{(e)}(\ell)$ is computed separately for each example, the resulting layer-wise profiles vary across examples.

The modality-specific layer-wise profiles derived from audio, visual, and joint audio-visual sinks are then combined according to the perceptual information required by the text prompt. For example, some prompts primarily require audio evidence, others visual evidence, while some require both modalities. Following prior work on modality assessment (Luo et al., 2026; Chung et al., 2026), we use the same multimodal LLM being evaluated with a lightweight modality-assessment prompt to obtain logits $(z_a^{(e)}, z_v^{(e)}, z_{av}^{(e)})$ for audio, visual, and joint audio-visual reasoning. We convert these logits into routing weights $(w_a^{(e)}, w_v^{(e)}, w_{av}^{(e)}) = \mathrm{softmax}(z_a^{(e)}, z_v^{(e)}, z_{av}^{(e)})$. The final reweighting strength at layer ℓ is then

$$\gamma_\ell^{(e)} = \gamma_{\mathrm{base}} \left(w_a^{(e)} \, \mathrm{shape}_a^{(e)}(\ell) + w_v^{(e)} \, \mathrm{shape}_v^{(e)}(\ell) + w_{av}^{(e)} \, \mathrm{shape}_{av}^{(e)}(\ell) \right). \tag{4}$$

The routing weights combine the three modality-specific profiles into a single layer-wise reweighting schedule. At each layer ℓ, the resulting $\gamma_\ell^{(e)}$ is applied to all perceptual sink positions in faithful heads through Eq. 2, rather than assigning different reweighting strengths to individual sink types. The schedule is computed once from the initial unmodified prefill pass. We then rerun the prefill with HARPS enabled to generate the first token and keep the same schedule fixed throughout subsequent autoregressive decoding. The only tuned hyperparameter is γ_{base}.

4 Experiments

4.1 Experimental Setup

Model Selection. We evaluate HARPS on four multimodal LLMs ranging from 3B to 30B parameters: Qwen2.5-Omni-3B and Qwen2.5-Omni-7B (Xu et al., 2025a), Qwen3-Omni-30B (MoE) (Xu et al., 2025b), and VideoLLaMA2-7B (Cheng et al., 2024). Together, these models cover multiple scales, both dense and mixture-of-experts architectures (Mu & Lin, 2025), and two distinct model families.

Evaluation Benchmarks. We evaluate HARPS from two complementary perspectives: hallucination mitigation and general multimodal capability.

- For hallucination, we use CMM (Leng et al., 2026) and open-ended audio-visual captioning on VGGSounder (Zverev et al., 2025), where hallucination is measured using $\mathrm{CHAIR_S}$ and $\mathrm{CHAIR_I}$ (Rohrbach et al., 2018).
- To assess whether the intervention preserves broader reasoning ability, we further evaluate on AV-SpeakerBench (Nguyen et al., 2026), WorldSense (Hong et al., 2026), Video-MME (Fu et al., 2025), MMAU (Sakshi et al., 2025), MME (Fu et al., 2026a), and OmniBench (Li et al., 2026), covering audio-visual, audio-only, visual-only, and image-audio understanding.

Table 1: **Main results across hallucination and general multimodal reasoning benchmarks**. Subscripts denote the change relative to the corresponding base model. **Bold** and underline indicate the best and second-best results within each model block.

	Video–Audio				Image–Audio	Audio-only	Visual-only
Method	**CMM**	**AV-Speaker**	**WorldSense**	**Video-MME**	**OmniBench**	**MMAU**	**MME**
Qwen2.5-Omni-3B	83.3	40.6	40.2	58.8	46.3	51.2	2238.6
+ *VCD* (Leng et al., 2024)	–	–	–	–	–	–	2230.6 $_{-8.0}$
+ *AVCD* (Jung et al., 2026a)	83.2 $_{-0.1}$	40.4 $_{-0.2}$	40.3 $_{+0.1}$	59.3 $_{+0.5}$	46.7 $_{+0.4}$	–	–
+ *MAD* (Chung et al., 2026)	84.0 $_{+0.7}$	41.2 $_{+0.6}$	40.2 $_{\pm0.0}$	59.4 $_{+0.6}$	35.2 $_{-11.1}$	–	–
+ *ASD* (Jung et al., 2026b)	83.7 $_{+0.4}$	40.5 $_{-0.1}$	40.4 $_{+0.2}$	58.7 $_{-0.1}$	46.4 $_{+0.1}$	–	–
+ *Ours*	**84.2** $_{+0.}$	**41.8** $_{+1.2}$	**41.2** $_{+1.0}$	**60.8** $_{+2.0}$	**47.5** $_{+1.2}$	**52.7** $_{+1.5}$	**2285.4** $_{+46.8}$
Qwen2.5-Omni-7B	85.5	42.6	40.8	62.9	51.4	44.9	2156.5
+ *VCD* (Leng et al., 2024)	–	–	–	–	–	–	2169.9 $_{+13.4}$
+ *AVCD* (Jung et al., 2026a)	85.5 $_{\pm0.0}$	42.7 $_{+0.1}$	41.4 $_{+0.6}$	63.3 $_{+0.4}$	51.5 $_{+0.1}$	–	–
+ *MAD* (Chung et al., 2026)	86.7 $_{+1.2}$	37.5 $_{-5.1}$	41.2 $_{+0.4}$	62.6 $_{-0.3}$	43.7 $_{-7.7}$	–	–
+ *ASD* (Jung et al., 2026b)	85.6 $_{+0.1}$	42.7 $_{+0.1}$	41.5 $_{+0.7}$	63.2 $_{+0.3}$	51.7 $_{+0.3}$	–	–
+ *Ours*	**87.1** $_{+1.}$	**43.5** $_{+0.9}$	**42.1** $_{+1.3}$	**63.6** $_{+0.7}$	**52.7** $_{+1.3}$	**47.9** $_{+3.0}$	**2182.9** $_{+26.4}$
VideoLLaMA2-7B	77.4	38.9	27.4	51.4	35.4	14.0	1805.3
+ *VCD* (Leng et al., 2024)	–	–	–	–	–	–	1726.0 $_{-79.3}$
+ *AVCD* (Jung et al., 2026a)	77.0 $_{-0.4}$	39.0 $_{+0.1}$	27.6 $_{+0.2}$	51.7 $_{+0.3}$	35.7 $_{+0.3}$	–	–
+ *MAD* (Chung et al., 2026)	77.6 $_{+0.2}$	39.1 $_{+0.2}$	27.8 $_{+0.4}$	51.8 $_{+0.4}$	35.5 $_{+0.1}$	–	–
+ *ASD* (Jung et al., 2026b)	77.9 $_{+0.5}$	39.1 $_{+0.2}$	27.5 $_{+0.1}$	51.8 $_{+0.4}$	35.6 $_{+0.2}$	–	–
+ *Ours*	**78.3** $_{+0.}$	**39.7** $_{+0.8}$	**28.2** $_{+0.8}$	**52.3** $_{+0.9}$	**36.2** $_{+0.8}$	**20.8** $_{+6.8}$	**1858.4** $_{+53.1}$
Qwen3-Omni-30B	87.2	51.9	43.3	63.5	58.5	72.1	2480.9
+ *VCD* (Leng et al., 2024)	–	–	–	–	–	–	2455.6 $_{-25.3}$
+ *AVCD* (Jung et al., 2026a)	87.2 $_{\pm0.0}$	52.1 $_{+0.2}$	43.6 $_{+0.3}$	64.0 $_{+0.5}$	58.4 $_{-0.1}$	–	–
+ *MAD* (Chung et al., 2026)	88.5 $_{+1.3}$	53.0 $_{+1.1}$	43.3 $_{\pm0.0}$	63.9 $_{+0.4}$	57.4 $_{-1.1}$	–	–
+ *ASD* (Jung et al., 2026b)	87.6 $_{+0.4}$	51.9 $_{\pm0.0}$	43.1 $_{-0.2}$	63.8 $_{+0.3}$	58.5 $_{\pm0.0}$	–	–
+ *Ours*	**88.6** $_{+1.}$	**53.1** $_{+1.2}$	**44.1** $_{+0.8}$	**64.6** $_{+1.1}$	**59.0** $_{+0.5}$	**73.8** $_{+1.7}$	**2488.5** $_{+7.6}$

Baselines. We compare HARPS against greedy decoding and several representative training-free hallucination mitigation methods. These include contrastive decoding approaches, namely VCD (Leng et al., 2024), AVCD (Jung et al., 2026a), and MAD (Chung et al., 2026), as well as ASD (Jung et al., 2026b), which directly modifies attention during inference. Detailed decoding and inference configurations are provided in Appendix E.

Evaluation Protocol. HARPS introduces a single tunable scalar, γ_{base}. Following prior training-free hallucination mitigation methods (Jung et al., 2026a; Chung et al., 2026; Jung et al., 2026b), we use AVHBench (Sung-Bin et al., 2025) as a development set and select γ_{base} on a fixed random subset of 300 examples for each model. **The selected value (Appendix G) is then kept unchanged across all subsequent evaluations** for that model, including both hallucination-focused and general multimodal reasoning benchmarks. Since AVHBench is used for hyperparameter selection, we exclude it from our main evaluation; for completeness, we report results on AVHBench in Appendix F, where HARPS also outperforms the compared training-free methods. For benchmark-based reasoning tasks, we follow the official evaluation protocols and report accuracy. For video and audio-video benchmarks, we evaluate the short- and medium-length subsets to reduce confounding from long-context frame-budget limitations (Tang et al., 2025). Evaluation prompts and additional implementation details are provided in Appendix D.

4.2 Main Results

HARPS consistently improves multimodal performance without capability trade-offs. As shown in Table 1, HARPS consistently improves both hallucination-focused and general multimodal reasoning benchmarks across all evaluated models, achieving the strongest overall performance among the compared methods. Among the 24 accuracy-based model–benchmark settings, HARPS improves the corresponding base model by at least 0.8 percentage points in 22 of them, with the highest gain of 6.8 percentage points. In contrast, each baseline exhibits at least one setting where performance falls below that of the corresponding base model. Together, these results show that HARPS delivers broad and consistent gains without compromising the models' existing multimodal capabilities, while incurring only little additional inference overhead (Sec. 4.3).

Table 2: **HARPS consistently reduces hallucination in open-ended audio-visual captioning.** CHAIR_S and CHAIR_I are reported on VGGSounder, with lower values indicating less hallucination. **Bold** and underline indicate the best and second-best results within each model block.

	Qwen2.5-Omni-3B		Qwen2.5-Omni-7B		Qwen3-Omni-30B		VideoLLaMA2-7B	
Method	$\text{CHAIR}_\text{S}\downarrow$	$\text{CHAIR}_\text{I}\downarrow$	$\text{CHAIR}_\text{S}\downarrow$	$\text{CHAIR}_\text{I}\downarrow$	$\text{CHAIR}_\text{S}\downarrow$	$\text{CHAIR}_\text{I}\downarrow$	$\text{CHAIR}_\text{S}\downarrow$	$\text{CHAIR}_\text{I}\downarrow$
Greedy decoding	71.08	39.66	67.25	48.83	75.25	29.35	29.83	24.98
AVCD (Jung et al., 2026a)	97.75	47.69	75.17	**34.03**	72.75	29.92	31.58	25.73
MAD (Chung et al., 2026)	71.17	48.56	82.17	49.35	**65.25**	34.52	29.92	24.89
ASD (Jung et al., 2026b)	71.50	40.21	67.33	47.50	75.25	29.35	29.92	**22.41**
Ours	**69.75**	**38.58**	**66.25**	43.71	74.17	**26.28**	**28.83**	23.73

HARPS generalizes across perceptual inputs, model families, and scales. As shown in Table 1, the gains of HARPS extend across audio-only, visual-only, image-audio, and audio-video settings. They also hold across both Qwen-Omni and VideoLLaMA2, and from 3B to 30B parameters, covering different model families, architectures, and scales. These results show that the effectiveness of HARPS is not tied to a particular perceptual input configuration or model design.

HARPS effectively reduces multimodal hallucination. On open-ended audio-visual captioning with VGGSounder (Table 2), HARPS lowers both CHAIR_S and CHAIR_I relative to greedy decoding for all four models, achieving the best result in six of the eight model–metric settings. Each compared baseline, in contrast, underperforms greedy decoding in at least one CHAIR setting. The same trend is observed on CMM in Table 1, where HARPS improves all four base models by 0.9–1.6 percentage points and achieves the best performance in every case.

4.3 Analysis

HARPS adds little inference overhead. HARPS requires only an initial unmodified prefill pass to construct the example-specific reweighting schedule, after which the schedule remains fixed throughout autoregressive decoding. As shown in Fig. 4, this results in only $1.11\times$ the latency of greedy decoding on Qwen3-Omni-30B. In comparison, ASD incurs $1.57\times$ latency, while the contrastive decoding methods AVCD and MAD require $2.62\times$ and $2.43\times$, respectively. Thus, HARPS retains the efficiency of a lightweight inference-time intervention while achieving substantially lower overhead than the compared training-free methods.

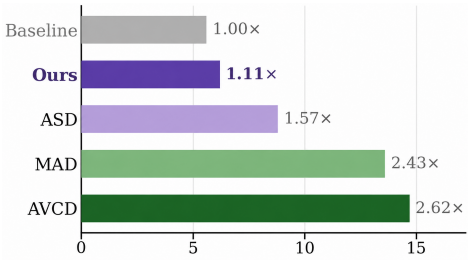


Figure 4: Inference time per question (s) on AVHBench with Qwen3-Omni-30B.

Intervening on faithful heads is most effective. A key design choice in HARPS is to apply the intervention only to faithful heads. We compare this choice with several alternatives on Qwen2.5-Omni-7B using the AVHBench development set. As shown in Table 3, suppressing or boosting unfaithful heads provides little benefit, while intervening on all heads or a random subset performs better. Applying the intervention to faithful heads achieves the strongest result among the fixed-γ variants, improving accuracy from 73.00 to 77.67.

Example- and prompt-specific adaptation further improves performance. Having established where to intervene, we next study how the reweighting strength should vary across layers and examples. Starting from faithful-head intervention with a fixed γ, Table 4 compares progressively more adaptive variants. A fixed layer-wise schedule does not improve over the constant baseline, yielding 77.33 versus 77.67. In contrast, using example-specific layer profiles together with prompt-dependent routing increases accuracy to 78.33. This shows that the gain from adaptive reweighting comes from tailoring the schedule to both the input and the perceptual requirements of the text prompt.

Qualitative analysis. Fig. 5 shows representative audio-visual reasoning examples comparing HARPS with the base model and other training-free methods. In the first example, several base-

Table 3: **Head-selection ablation.** Accuracy on the AVHBench development set with Qwen2.5-Omni-7B.

Intervention	Accuracy
No intervention	73.00
Suppress unfaithful heads	73.00
Boost sinks in unfaithful heads	73.33
Boost sinks in all heads	76.00
Boost sinks in random heads	77.11
Boost sinks in faithful heads	**77.67**

Table 4: **Adaptive-reweighting ablation.** Accuracy on the AVHBench development set with Qwen2.5-Omni-7B.

Reweighting strategy	Accuracy
No intervention	73.00
Constant γ	77.67
+ layer-wise schedule	77.33
+ example/prompt adaptation	**78.33**

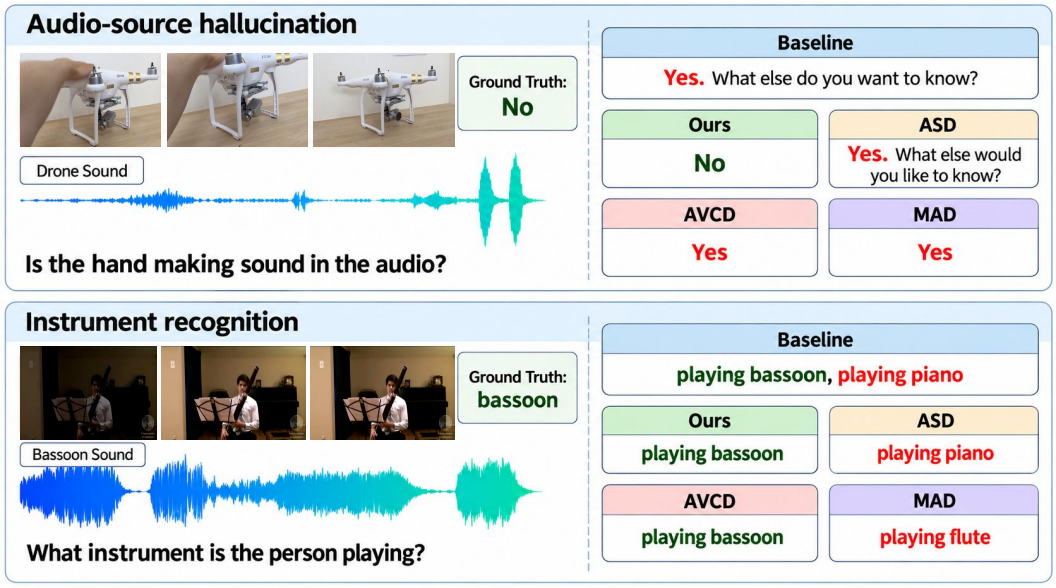


Figure 5: **Qualitative comparison on audio-visual reasoning examples.** In the top example, the audio contains a drone sound, but the question asks whether the hand is producing it; HARPS correctly answers "No," while the base model and several training-free baselines incorrectly attribute the sound to the hand. In the bottom example, multiple instruments are visible, but only the bassoon is heard in the audio; HARPS correctly identifies the sounding instrument, whereas several baselines are distracted by visually present alternatives.

lines correctly detect that a sound is present but incorrectly attribute it to the visible hand, whereas HARPS identifies that the hand is not the sound source. In the second example, multiple instruments are visible, requiring the model to use the audio to determine which instrument is actually sounding; HARPS correctly identifies the bassoon while several baselines select visually plausible alternatives. These examples illustrate how HARPS better grounds its predictions in the perceptual evidence required by the text prompt.

5 Conclusion

In this work, we study multimodal hallucination in multimodal LLMs and introduce HARPS, a training-free inference-time method for hallucination mitigation. Motivated by our analysis showing that faithful and unfaithful attention heads interact differently with perceptual sink tokens, HARPS selectively reweights attention to these tokens in faithful heads and further adapts the intervention across decoder layers and input-prompt pairs. Experiments across multiple multimodal LLMs and a diverse set of hallucination and general reasoning benchmarks show consistent improvements over the underlying models and existing training-free methods, with low inference overhead. Overall, our results demonstrate the effectiveness of selective and adaptive attention reweighting for mitigating multimodal hallucination.

AI USE STATEMENT

In this work, we used generative AI tools to assist with method implementation. We have not used generative AI tools to develop research ideas or hypotheses, design or provide feedback on research methodology or experiments, interpret results, generate synthetic datasets, or support qualitative or thematic data analysis; tasks involving theoretical models, mathematical claims or proofs, translation, and dataset cleaning or reformatting are not applicable to this work. Additionally, we used generative AI tools to create and refine scientific figures, edit and debug software code, draft and revise portions of the manuscript, improve manuscript readability, suggest manuscript structure, and propose paper titles. We have reviewed all AI-assisted work. All AI-assisted code, figures, and manuscript content were reviewed, revised, and finalized by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies hallucination mitigation in multimodal large language models using publicly available datasets and pretrained models. We do not collect new human-subject data or use sensitive personal information. Our experiments are intended to improve the reliability and grounding of multimodal models, and we are not aware of additional ethical risks beyond those generally associated with the development and deployment of large language models. The authors have reviewed the study for compliance with the ICLR Code of Ethics and take responsibility for the conduct and reporting of the research.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental details needed to reproduce our results throughout the main paper and appendix. Section 3 specifies the complete HARPS procedure, including the intervention-based identification of faithful and unfaithful attention heads, perceptual sink-token detection and head-selective reweighting, and the example- and prompt-adaptive layer-wise reweighting schedule. Section 4.1 describes the evaluated models, benchmarks, baselines, hyperparameter-selection protocol, and evaluation settings. In particular, the only tuned intervention hyperparameter, γ_{base} is selected separately for each model on a fixed 300-example subset of AVHBench and then kept unchanged across subsequent evaluations for that model. Appendix A details the deterministic construction of grounded and hallucinated token sets and the head-attribution procedure; Appendix B reports the model-specific sink-detection thresholds and sink dimensions; Appendix C provides the modality-assessment prompt and routing-weight computation; Appendices D and E specify the exact benchmark prompts and inference configurations; and Appendix G reports the ablations and hyperparameter sweep of γ_{base} used to motivate the final design choices. Together, these specifications provide the prompts, model-specific parameters, intervention rules, evaluation protocol, and hyperparameter settings required to reproduce the reported experiments. We will release the code upon acceptance.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 961–970, 2015.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, volume 2024, pp. 54158–54183, 2024.

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

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28652–28668, 2026.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 11198–11201, 2024.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, June 2025.

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, 2026a.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long MA, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *Advances in Neural Information Processing Systems*, 38:75300–75320, 2026b.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Advances in Neural Information Processing Systems*, 38:41819–41886, 2026.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26574–26585. IEEE, 2024.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. In *International Conference on Learning Representations*, volume 2026, pp. 52423–52443, 2026.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427. IEEE, 2024.

Chaeyoung Jung, Youngjoon Jang, and Joon Son Chung. Avcd: Mitigating hallucinations in audio-visual large language models through contrastive decoding. *Advances in Neural Information Processing Systems*, 38:63143–63174, 2026a.

Jihoo Jung, Chaeyoung Jung, Ji-Hoon Kim, and Joon Son Chung. Probing cross-modal information hubs in audio-visual LLMs. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=nxKRwB1J63.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations*, volume 2025, pp. 87676–87703, 2025.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25125–25148. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/kong24a.html.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 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. *Advances in Neural Information Processing Systems*, 38, 2026.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023a.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023b.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Hangyu Guo, Yiming Liang, Jiaheng Liu, Noah Wang, Jian Yang, Siwei Wu, et al. Omnibench: Towards the future of universal omni-language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Jiayun Luo, Wan-Cyuan Chris Fan, Lyuyang Wang, Xiangteng He, Tanzila Rahman, Purang Abolmaesumi, and Leonid Sigal. To sink or not to sink: Visual information pathways in large vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 93626–93660, 2026.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32, 2019.

Siyuan Mu and Sen Lin. A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. *arXiv preprint arXiv:2503.07137*, 2025.

KiHyun Nam, Jungwoo Heo, Siu Bae, Ha-Jin Yu, and Joon Son Chung. Speakerllm: A speaker-specialized audio-llm for speaker understanding and verification reasoning. *arXiv preprint arXiv:2605.15044*, 2026.

Le Thien Phuc Nguyen, Zhuoran Yu, Samuel Low Yu Hang, Subin An, Jeongik Lee, Yohan Ban, SeungEun Chung, Thanh-Huy Nguyen, JuWan Maeng, Soochahn Lee, et al. See, hear, and understand: Benchmarking audiovisual human speech understanding in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2272–2283, 2026.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Yuqi Pang, Bowen Yang, Haoqin Tu, Yun Cao, and Zeyu Zhang. Language models can see better: Visual contrastive decoding for llm multimodal reasoning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

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.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 84929–84964, 2025.

Ramaneswaran Selvakumar, Kaousheik Jayakumar, S Sakshi, Sreyan Ghosh, Ruohan Gao, and Dinesh Manocha. Do audio-visual large language models really see and hear? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 5892–5902, June 2026.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=F7aAhfitX6.

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.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29118–29128. IEEE, 2025.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the association for computational linguistics: ACL 2024*, pp. 15840–15853, 2024b.

Chang-Hsun Wu, Kai-Po Chang, Yu-Yang Sheng, Hung-Kai Chung, Kuei-Chun Wang, and Yu-Chiang Frank Wang. Season: Mitigating temporal hallucination in video large language models via self-diagnostic contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11096–11105, 2026.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 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.

Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *European conference on computer vision*, pp. 603–619. Springer, 2018.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *International Conference on Learning Representations*, volume 2025, pp. 51546–51568, 2025a.

Yudong Yang, Jimin Zhuang, Guangzhi Sun, Changli Tang, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, Zejun Ma, and Chao Zhang. Audio-centric video understanding benchmark without text shortcut. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6580–6598, 2025b.

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

Suho Yoo, Youngjoon Jang, and Joon Son Chung. On the nature of attention sink that shapes decoding strategy in omni-llms. *arXiv preprint arXiv:2603.14337*, 2026.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9637–9662, 2024.

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.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Chaochen Gu, Xiaosong Yuan, Shaotian Yan, Jiawei Cao, Hao Cheng, Kaijie Wu, and Jieping Ye. Shallow focus, deep fixes: Enhancing shallow layers vision attention sinks to alleviate hallucination in LVLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3512–3534, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.174. URL https://aclanthology.org/2025.emnlp-main.174/.

Daniil Zverev, Thaddäus Wiedemer, Ameya Prabhu, Matthias Bethge, Wieland Brendel, and A Koepke. Vggsounder: Audio-visual evaluations for foundation models. In *Proceedings of the ieee/cvf international conference on computer vision*, pp. 1027–1037, 2025.

A DETERMINISTIC DETERMINATION OF HALLUCINATED- AND GROUNDED-TOKEN SETS

A.1 CONSTRAINED GENERATION FOR HALLUCINATION HEADS DETECTION

The contrastive score of equation 1 requires partitioning the model's own output into hallucinated and grounded content. In free-form captioning this partition is itself a judgement: entity mentions must be extracted from prose, resolved against a ground-truth annotation, and adjudicated where the surface forms disagree, whether by a matching heuristic or by a second model. Either route introduces error that is difficult to separate from the attribution it feeds.

We avoid the judgement entirely by constraining the output. Each corpus is annotated over a closed label set $\mathcal{L}$—the AudioSet ontology, the ActivityNet activity classes, and so on—and we prompt the model to answer from that set alone:

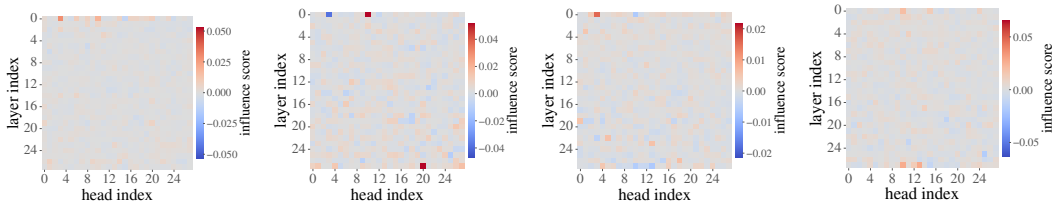
```
Describe what you {hear | see} in the {audio | video}.
Answer with a comma-separated list drawn only from the following
labels:
{ℓ₁, ℓ₂, ..., ℓ_|L|}
```

with *hear*/*audio* used for AudioSet and LibriSpeech and *see*/*video* for ActivityNet and YouTube-VOS. The response is a list of labels, and each element is either in the clip's ground-truth annotation $\mathcal{G}$ or is not. Writing y for the generated sequence and $\mathcal{E}(y)$ for the token positions spanning its list elements,

$$\mathcal{T}_{\text{hal}}(y) = \{\, t \in \mathcal{E}(y) : \ell(t) \notin \mathcal{G} \,\}, \qquad \mathcal{T}_{\text{gnd}}(y) = \{\, t \in \mathcal{E}(y) : \ell(t) \in \mathcal{G} \,\}, \tag{5}$$

where $\ell(t)$ is the label whose span contains position t. Separators and any preamble fall in neither set. The partition is a set membership test against the corpus annotation, with no parser, no matching heuristic, and no judge model, so it is reproducible from the generation and the annotation alone.

A.2 PER-CORPUS ATTRIBUTION



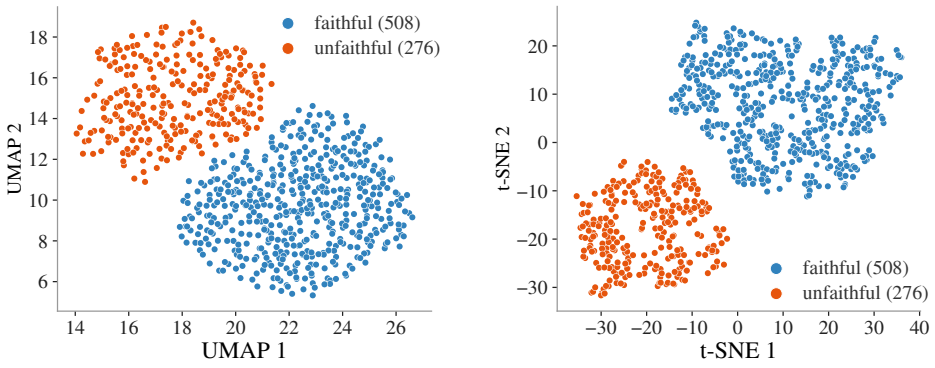
(a) ActivityNet (b) YouTube-VOS (c) LibriSpeech (d) AudioSet

Figure 6: Contrastive influence $s_h^{(p)}$ over the 28×28 head grid of Qwen2.5-Omni-7B, for each of the four detection corpora.

Figure 6 shows $s_h^{(p)}$ for all four corpora. The score distributions are diffuse, with most heads lying near zero and no contiguous region dominating across the head grid. The heads with larger-magnitude scores also vary across corpora, indicating that hallucination-associated behavior is not concentrated in a single fixed subset of heads across all content types. Score magnitudes further differ across corpora, with LibriSpeech exhibiting the smallest overall variation.

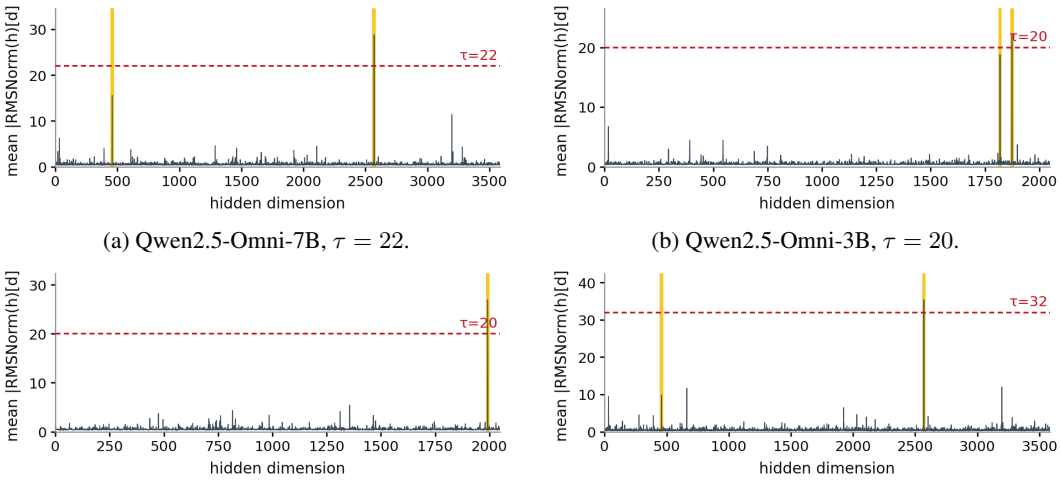
A.3 TWO-DIMENSIONAL VIEW OF THE HEAD TAXONOMY

Collecting the four scores gives each head a coordinate in $\mathbb{R}^4$. Figure 7 projects this to two dimensions under UMAP (McInnes et al., 2018) and t-SNE (Van der Maaten & Hinton, 2008), with the two classes coloured. We include the figure to show that the two groups differ across the whole signature rather than along any single corpus, which the per-panel view of Figure 6 does not make apparent.



(a) UMAP (b) t-SNE

Figure 7: Heads embedded from their four-corpus signature $\left(s_h^{(1)}, \ldots, s_h^{(4)}\right)$ and coloured by the partition on Qwen2.5-Omni-7B.



(a) Qwen2.5-Omni-7B, $\tau = 22$. (b) Qwen2.5-Omni-3B, $\tau = 20$.

(c) Qwen3-Omni-30B, $\tau = 20$. (d) VideoLLaMA2-7B, $\tau = 32$.

Figure 8: Selection of the sink-dimension threshold τ for each backbone. We plot the mean absolute RMS-normalized activation for every hidden dimension. A small number of dimensions exhibit massive activations that are clearly separated from the remaining dimensions. The red dashed line denotes the model-specific threshold used to identify sink dimensions, while highlighted dimensions exceed the selected threshold.

B Sink Tokens

We determine τ directly from the activation distributions shown in Fig. 8. Across all four models, most hidden dimensions have relatively small activation magnitudes, while a small number of dimensions exhibit clearly separated massive activations. We therefore choose τ at the boundary separating these outlier dimensions from the ordinary activation regime. This yields $\tau = 22$ for Qwen2.5-Omni-7B, $\tau = 20$ for Qwen2.5-Omni-3B, $\tau = 20$ for Qwen3-Omni-30B, and $\tau = 32$ for VideoLLaMA2-7B. The selected thresholds are fixed for each backbone and used across all samples and evaluation benchmarks. Table 5 reports the parameters used for each backbone.

C Modality Assessment

Our intervention is conditioned on the perceptual information required by the text prompt. Rather than applying the same layer-wise profile to every example, we assess whether the prompt relies

Table 5: Model-specific parameters used for sink-token identification. τ denotes the sink activation threshold.

Model	τ	D_{sink}
Qwen2.5-Omni-7B	22	(458, 2570)
Qwen2.5-Omni-3B	20	(1819, 1874)
Qwen3-Omni-30B	20	(1992)
VideoLLaMA2-7B	32	(458, 2570)

primarily on *Audio*, *Visual*, or joint *Audio–Visual* reasoning. The resulting scores are converted into soft routing weights used to combine the corresponding modality-specific layer-wise profiles.

Specifically, given a text prompt q, we query the same multimodal LLM used for generation with the following modality-assessment prompt:

```
You are classifying questions by the modality they ask about.  Audio
if the question is about sounds, hearing, or what is audible.  Visual
if the question is about images, objects, or what is visible.  AV if
answering requires comparing or matching what is heard against what
is seen (cross-modal consistency, joint description).  Respond with
exactly one word:  Audio, Visual, or AV.
Question:  {q}
Classification:
```

Although the prompt constrains the generated answer to a single modality label, we do not use the generated label itself for routing. Instead, at the first output position following `Classification:`, we extract the pre-softmax logits associated with the three candidate labels `Audio`, `Visual`, and `AV`. Let these logits be

$$\mathbf{z}(q) = [z_A(q),\, z_V(q),\, z_{AV}(q)]\,. \tag{6}$$

We convert them into normalized routing weights using a softmax:

$$\mathbf{w}(q) = \mathrm{softmax}(\mathbf{z}(q))\,, \tag{7}$$

where $\mathbf{w}(q) = [w_A, w_V, w_{AV}]$ and $w_A + w_V + w_{AV} = 1$.

The routing weights reflect the perceptual requirements of the text prompt rather than simply the modalities present in the input.

D Prompt configuration

Every benchmark is run with its official prompt wherever the model family supports it. Unless otherwise specified, all methods use the same prompt as the corresponding baseline on that model, so within-model comparisons remain controlled with respect to the instruction. Method-specific prompt differences and cross-family prompt differences are documented below, with the latter summarized in Sec. D.2.

D.1 Qwen2.5-Omni and Qwen3-Omni

All Qwen runs use the model's own persona string as the system prompt, with the single exception of OmniBench:

```
You are Qwen, a virtual human developed by the Qwen Team, Alibaba
Group, capable of perceiving auditory and visual inputs, as well as
generating text and speech.
```

AVHBench. The question is followed by a fixed suffix constraining the answer to a binary decision.

```
{question} Answer with only 'Yes' or 'No'.
```

MAD is run with the suffix given in its reference implementation, "`{question} Answer only 'Yes' or 'No'. Do not include any explanation.`", rather than with ours, so that the method is evaluated as its authors specify.

CMM. The question is followed by a fixed suffix constraining the answer to a binary decision.

```
{question} Answer with yes or no.
```

WorldSense. We use the official VLMEvalKit (Duan et al., 2024) prompt verbatim, assembled in the order system prompt, video, audio, template, question.

```
Carefully watch this video and pay attention to every detail. Based
on your observations, select the best option that accurately addresses
the question.
[video] [audio]
These are the frames of a video and the corresponding audio. Select
the best answer to the following multiple-choice question based on the
video. Respond with only the letter (A, B, C, or D) of the correct
option.
Question: {question}
{candidates}
Answer:
```

Video-MME. Video-MME is evaluated through the WorldSense backend (Duan et al., 2024) and therefore uses the identical prompt and sampling configuration.

AV-SpeakerBench. We use the original paper's template.

```
Select the best answer to the following multiple-choice question based
on the video. Respond with only the letter (A, B, C, or D) of the
correct option.
{question}
{choices}
The best answer is:
```

OmniBench. OmniBench is the one benchmark for which we depart from the persona system prompt, using the neutral "`You are a helpful assistant.`" instead.

```
Answer the following multiple-choice question using the image and the
audio.
{question}
{options}
Respond with only the letter (A, B, C, or D) of the correct option.
Do not explain.
```

D.2 VIDEOLLAMA2

VideoLLaMA2 is run without a system prompt, using the plain chat template. The binary benchmarks use the same suffixes as the Qwen runs, so those numbers are directly comparable across families.

Table 6: Inference configuration

Model	Params	Backbone	FPS	Max Frames
Qwen2.5-Omni	3B	Dense	1	20
Qwen2.5-Omni	7B	Dense	1	20
Qwen3-Omni	30B	MoE	1	20
VideoLLaMA2	7B	Dense	1	20

Table 7: **Schedule shape.** Boost on the 508-faithful heads core, AVHBench. Reverse weights layers where that class of sink is *sparse*; forward is the same schedule inverted; flat applies one γ everywhere. Best γ_{base} per row, selected on DEV.

Schedule	DEV
Flat	77.67
Forward (dense-first)	76.67
Reverse (sparse-first)	77.33
Reverse, example-specific **(ours)**	**78.33**

AVHBench

```
{question} Answer with only 'Yes' or 'No'.
```

CMM

```
{question} Answer with yes or no.
```

The four multiple-choice benchmarks—AV-SpeakerBench, OmniBench, Video-MME, and WorldSense—are instead routed through a single shared template:

```
{question}
A. {option_a}
B. {option_b}
...
Answer with the option's letter from the given choices directly.
```

Caveat. VideoLLaMA2 therefore does not use the official WorldSense, Video-MME, or AV-SpeakerBench prompts that the Qwen runs use. We standardized these because the model answers descriptively otherwise, which the official templates do not suppress. The shared template is applied identically across all five methods, so comparisons *within* VideoLLaMA2 remain controlled and the deltas reported in Table 1 are unaffected. Absolute VideoLLaMA2 numbers on those three benchmarks are not strictly prompt-matched to the Qwen numbers and should not be read as a cross-family comparison.

E Inference configuration.

All models are run with greedy decoding, and all remaining parameters follow each model's official inference code (Table 6). Videos are sampled at 1 fps for Qwen2.5-Omni (Xu et al., 2025a), VideoLLaMA2 (Cheng et al., 2024), and Qwen3-Omni (Xu et al., 2025b), matching each repository's default. Prompt configurations are provided in Appendix D.

F AVHBench Performance

AVHBench is where every training-free method for audio-visual hallucination is tuned, including ours, so we report it here rather than in Table 1 (Sec. 4.1). Figure 9 gives overall accuracy for readers

Table 8: **Proxy datasets for the head taxonomy.** The head set is derived by zero-ablation on an audio probe × a visual probe; FULL AVHBench accuracy (%). AS: AudioSet, LS: LibriSpeech, AN: ActivityNet, YT: YouTube-VOS. The last row retains only heads that remain faithful across all four proxy corpora. For this experiment, we use a fixed γ for reweighting.

Audio × visual probe	DEV
LS × YT	75.73
AS × YT	76.29
AS × AN	76.39
LS × AN	76.74
Always-faithful core (508 heads)	**77.20**

Table 9: **Boost strength γ_{base}.** AVHBench DEV accuracy (%), n=300, using the example-specific schedule at each model's default sink threshold. Shaded is the reported configuration.

(a) Qwen2.5-Omni-7B (τ=22)

γ_{base}	base	1	2	3	4	5	6	8
acc.	73.00	75.67	76.67	76.67	77.67	76.67	77.00	**78.33**

(b) Qwen2.5-Omni-3B (τ=20)

γ_{base}	base	3	5	8
acc.	71.67	73.67	**74.33**	73.00

(c) Qwen3-Omni-30B (τ=20)

γ_{base}	base	2	3	4	5
acc.	75.00	76.67	77.00	**77.33**	75.33

(d) VideoLLaMA2 (τ=32)

γ_{base}	base	1	2	3
acc.	69.33	70.00	**71.33**	70.33

who want the comparison against published numbers. The overall pattern is consistent with the held-out evaluations: our method achieves the best performance on Qwen2.5-Omni at both scales and on VideoLLaMA2, and ties with MAD on Qwen3-Omni. The margins are smaller than in Table 1 and should be read with the caveat above; a method that has been fitted to a benchmark will do better on it than its held-out behaviour warrants, by an amount that differs per method and cannot be recovered after the fact.

G Ablation Studies

We conduct all ablation studies with Qwen2.5-Omni-7B. We randomly select 300 questions from AVHBench as a development set and use this subset for hyperparameter selection and design ablations. Unless otherwise specified, all results in this section report accuracy on this development set. We study three aspects of our intervention: (i) the depth-dependent shape of the reweighting schedule, (ii) the choice of proxy datasets used to identify faithful heads, and (iii) the overall intervention strength.

Reweighting schedule. We first study how the reweighting strength should vary across decoder depth. Table 7 compares a flat schedule, which applies the same γ at every layer, with two depth-dependent schedules. The *forward* schedule assigns larger weights to layers where the corresponding sink type is more prevalent, while the *reverse* schedule emphasizes layers where it is relatively sparse. A fixed reverse schedule does not consistently improve over a flat intervention (77.33 vs. 77.67), indicating that a single global depth profile is insufficient across inputs. In contrast, adapting the schedule for each example achieves the best performance, reaching 78.33% accuracy. This supports our use of input-dependent reweighting rather than applying a fixed depth schedule to every example.

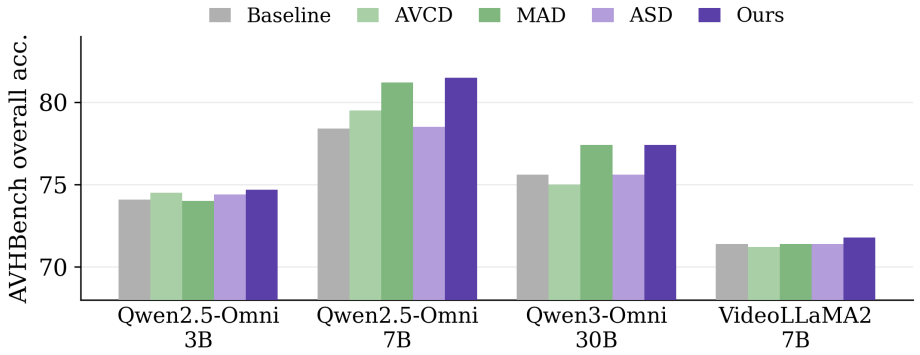


Figure 9: AVHBench overall accuracy. Green is contrastive decoding, purple attention manipulation. Note that γ_{base} is tuned on AVHBench, as are the hyperparameters of all three baselines, so these numbers are not a clean held-out comparison.

Robustness of the head taxonomy. We next study whether the faithful-head set should be identified from a single pair of audio and visual proxy corpora or from their intersection across all four corpora. Table 8 compares the four possible audio–visual proxy pairs while keeping the intervention fixed. Using any single pair yields similar AVHBench accuracy, ranging from 75.73% to 76.74%. In contrast, retaining only heads that remain faithful across all four corpora yields the strongest result at 77.20%. We therefore use this always-faithful intersection for the main method.

Intervention strength across models. We next examine the sensitivity to the global boost scale γ_{base} across four model backbones. Table 9 reports AVHBench development accuracy on the same 300-question split, using each model's example-specific schedule and default sink threshold. Overall, the method is robust to a range of boost strengths, but the optimal scale is model-dependent. For Qwen2.5-Omni-7B, performance increases from 73.00% to 78.33%, with the best result at $\gamma_{\text{base}} = 8$. Qwen2.5-Omni-3B improves from 71.67% to 74.33%, peaking at $\gamma_{\text{base}} = 5$. Qwen3-Omni-30B improves from 75.00% to 77.33% at $\gamma_{\text{base}} = 4$, and VideoLLaMA2 improves from 69.33% to 71.33% at $\gamma_{\text{base}} = 2$. In all four cases, overly large boosts can degrade performance, indicating that the intervention should amplify sink-mediated information without overwhelming the model's original attention distribution. We therefore tune γ_{base} per model on the development set and use the shaded configurations in the reported experiments.

H What Semantic Information Do Sink Tokens Carry?

Our intervention assumes that sink tokens are not merely attention artifacts, but retain modality-relevant information that can be selectively recovered by particular attention heads. We therefore analyze the semantic content carried by sink tokens and ask whether faithful and unfaithful heads read the same sink differently.

Head-conditioned OV-circuit decoding. Given a sink token j at decoder layer ℓ, we isolate its contribution through each attention head h. Let $v_j^{(\ell,h)}$ denote the value vector of the sink token for head h. We project this value through the head-specific output matrix and decode the resulting residual-stream contribution using the model's language-model head:

$$z_{j,h}^{(\ell)} = \text{lm_head}\left(\text{RMSNorm}\left(W_O^{(\ell,h)} v_j^{(\ell,h)}\right)\right). \tag{8}$$

We then take the top-K vocabulary entries under $z_{j,h}^{(\ell)}$ as a textual description of the information exposed by head h from sink token j. Importantly, this probe is query-independent: it considers only the sink's value pathway through the OV circuit, without using the attention weight assigned to the sink by a particular query token.

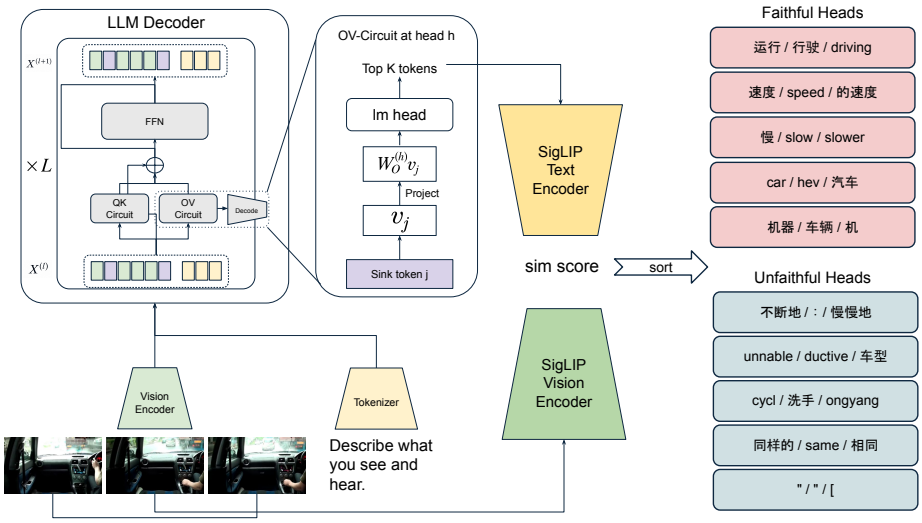


Figure 10: **Pipeline for probing the semantic content of sink tokens.** For a sink token j at layer ℓ, we pass its value vector through the OV circuit of each attention head h and decode the resulting residual-stream contribution with the model's LM head. The top-K vocabulary predictions provide a head-conditioned interpretation of the sink token. We measure their semantic alignment with the perceptual input using SigLIP for visual inputs and CLAP for audio inputs. For simplicity, the figure illustrates the visual/SigLIP branch; the audio/CLAP branch follows the same procedure.

Measuring semantic relevance. To determine whether the decoded tokens reflect the underlying perceptual content, we measure their semantic similarity to the corresponding input modality. For visual inputs, we use SigLIP (Zhai et al., 2023) to compare the decoded top-K tokens with the visual content; for audio inputs, we analogously use CLAP (Wu et al., 2023). These similarity scores are used only for analysis and do not affect head selection or our intervention.

Figure 10 illustrates the procedure for the visual branch using SigLIP; the audio branch follows the same procedure with CLAP. For a fixed sink token, faithful heads often expose recognizable concepts associated with the input, such as *driving*, *speed*, or *car* for a driving scene. In contrast, decoding the same sink through unfaithful heads frequently yields weakly related words, subword fragments, or punctuation. Thus, the semantic content recovered from a sink token is strongly head dependent.

Faithful versus unfaithful readers. We apply this analysis separately to the faithful and unfaithful head sets defined in Sec. 3.1. For each $(\text{clip}, \ell, \text{sink})$ instance, we decode the sink through every head of that layer and compare the resulting vocabulary predictions across the two groups. Faithful heads produce both more semantically aligned predictions and a larger fraction of interpretable words, whereas unfaithful heads are more likely to decode the same sink into unrelated or malformed tokens. These observations suggest that sink tokens themselves can preserve meaningful modality information, while the OV circuits of different heads determine whether that information is transformed into a useful direction in the residual stream.

This distinction motivates our intervention strategy: rather than uniformly increasing attention to perceptual sink tokens across all heads, we reweight attention to these tokens only in faithful heads, whose OV circuits are more likely to yield representations aligned with the underlying perceptual content.