RELEVANT EVIDENCE DECODING FOR AUDIO-VISUAL HALLUCINATION MITIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-Visual Large Language Models (AV-LLMs) remain prone to cross-modal hallucinations, where one modality incorrectly affects predictions about another. Although contrastive decoding reduces hallucinations in vision-language models, its direct extension to AV-LLMs overlooks a key challenge: different questions require different perceptual evidence, including audio, video, or their interaction. Notably, we observe that joint audio-visual inference can weaken the prediction even when a model can recover the correct answer from a single informative modality. For example, when asked which instrument is heard, a model may correctly predict violin from the audio alone. Once a video showing a guitar is added, its confidence in violin may drop. In this paper, we introduce **R**elevant **E**vidence **D**ecoding (**RED**), a training-free method that identifies question-relevant evidence and selectively strengthens its contribution. **RED** uses pointwise mutual information to quantify the predictive support provided by audio and video beyond the question alone. It decomposes their joint contribution into audio, video, and residual interaction components. A question-only inference pass determines the required evidence type, after which the model augments the original audio-visual prediction with the corresponding PMI contribution. Across three audio-visual hallucination benchmarks and three AV-LLMs, **RED** improves accuracy over standard decoding by up to 7% on CMM, 6.3% on AVHBench, and 3.8% on SVHalluc, with an average relative time to first token of 1.5× standard decoding.

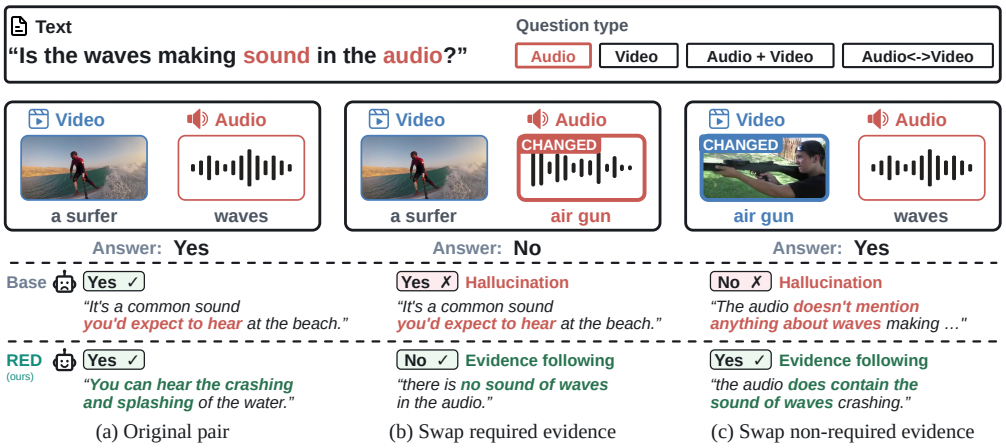
1 INTRODUCTION

Multimodal large language models (MLLMs) extend LLMs to perceptual inputs such as vision (Liu et al., 2023; Alayrac et al., 2022) and audio (Cheng et al., 2024; Xu et al., 2025), but additional modalities also introduce new sources of hallucination (Li et al., 2025; Zhao et al., 2026b). Prior work has linked these hallucinations to modality dominance (Leng et al., 2025; Ye et al., 2026), cross-modal interference (Li et al., 2026; Cai et al., 2025), and spurious inter-modality correlations (Leng et al., 2025). Consequently, more perceptual information does not necessarily yield more reliable predictions. In particular, we find that joint audio-visual inference can weaken a prediction even when the modality required by the question alone provides sufficient evidence for the correct answer.

Contrastive decoding (CD) has emerged as a promising approach to mitigate hallucinations in LLMs (Shi et al., 2024; Chuang et al., 2024) and VLMs (Leng et al., 2024; Wang et al., 2024; Kim et al., 2024). CD compares predictions from the same model under different conditioning inputs and uses their difference to adjust the original prediction without additional training. Recent methods have extended this idea to AV-LLMs by accounting for modality dominance (Jung et al., 2025), adaptively weighting modality-specific branches (Chung et al., 2026), or modeling cross-modal information flow (Jung et al., 2026). However, these methods focus on balancing or integrating multiple modalities rather than aligning the decoding correction with the evidence required by each question.

In practice, a correct prediction does not necessarily imply appropriate use of perceptual evidence (Xiao et al., 2024; Yin et al., 2025). A model may produce the correct answer from language priors (Favero et al., 2024), correlated cues, or perceptual evidence that is not required by the question (Baid et al., 2026). For example, an audio-dependent question can be answered correctly using

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.



(a) Original pair (b) Swap required evidence (c) Swap non-required evidence

Figure 1: **Motivation of this study.** (a) A correct prediction does not necessarily indicate successful hallucination mitigation. (b) When question-required evidence changes, the prediction should change accordingly. (c) When non-required evidence changes, it should remain unchanged. The base model violates both behaviors, whereas **RED** follows the evidence required by the question.

correlated visual evidence while the relevant audio evidence has little effect on the prediction, as shown in Figure 1(a). We argue that the required evidence should instead have a direct and selective influence on prediction: *changing question-required evidence should alter the prediction, whereas changing irrelevant evidence should have substantially less effect* (please see Figure 1(b) and (c)). The central challenge is to couple two distinct quantities: what evidence the question requires and how strongly that evidence supports each candidate prediction. This motivates evidence-following decoding, which aligns the correction with question-required predictive support rather than with any signal that happens to favor the answer.

In this study, we introduce **R**elevant **E**vidence **D**ecoding (**RED**), a training-free decoding method that identifies the evidence required by the question and aligns the decoding correction with the measured predictive support of that evidence. A question-only forward pass identifies the required evidence. Pointwise Mutual Information (PMI) (Church & Hanks, 1990) measures its predictive support beyond the question and language context by decomposing the joint audio-visual contribution into audio, video, and residual interaction components. **RED** refines the original audio-visual prediction using the PMI contribution corresponding to the required evidence. Contrary to existing CD methods that modulate modality-level branches (Jung et al., 2025; Chung et al., 2026) or amplify the overall contribution of additional conditioning information (Shi et al., 2024; Favero et al., 2024), **RED** directly encourages predictions to follow question-required evidence rather than question-irrelevant cues.

We evaluate our method with three backbones (*e.g.*, VideoLLaMA2.1-7B-AV (Cheng et al., 2024), Qwen2.5-Omni-3B, and -7B (Xu et al., 2025)) on three audio-visual hallucination benchmarks: CMM (Leng et al., 2025), AVHBench (Kim et al., 2025), and SVHalluc (Zhang et al., 2026). **RED** consistently outperforms existing decoding methods across all benchmarks and backbones. Results on two general audio-visual question answering benchmarks (Li et al., 2022; Kim & Seo, 2026) further show that its benefits transfer to general audio-visual reasoning. Extensive analyses show complementary gains from evidence selection and PMI decomposition, with RED's average relative time to first token (TTFT) at 1.5× standard decoding.

- We characterize evidence following through controlled modality interventions, showing that even correct AV-LLM predictions can respond weakly to question-required evidence and remain sensitive to non-required evidence.
- We propose **RED**, a training-free CD method that uses PMI to decompose audio-visual predictive changes relative to a common question-only reference and selectively strengthens the contribution corresponding to question-required evidence.
- Across three AV-LLMs and five benchmarks, **RED** generally outperforms existing decoding methods on both hallucination and general audio-visual reasoning tasks, while extensive analyses validate the effectiveness of evidence selection and PMI decomposition.

2 Related Work

Multimodal hallucination. Hallucination in vision-language models (VLMs) has been widely studied, particularly in relation to object hallucination and excessive reliance on language priors (Li et al., 2023; Guan et al., 2024). When models jointly process audio and video, additional failures can arise from modality dominance, cross-modal interference, and spurious inter-modality correlations (Kim et al., 2025; Leng et al., 2025; Zhang et al., 2026). Controlled perturbations and modality interventions reveal how predictions depend on individual input sources (Yin et al., 2025; Baid et al., 2026). However, input dependence alone does not indicate whether a prediction follows the evidence required by a particular question. Recent training-based work also promotes sensitivity to relevant modalities and invariance to irrelevant ones through modality-aware preference optimization (Chaubey et al., 2026). We study *evidence following* as a question-conditioned criterion that distinguishes what evidence is required from how it contributes to the prediction: predictions should respond to required evidence while remaining stable to non-required evidence.

Decoding-based hallucination mitigation. Contrastive decoding (CD) mitigates hallucination by adjusting predictions using differences between distributions obtained under alternative input conditions (Leng et al., 2024; Wang et al., 2024; Kim et al., 2024; Huo et al., 2025). Existing VLM methods construct these contrasts from distorted visual inputs, perturbed instructions, generated descriptions, or self-introspective signals (Leng et al., 2024; Wang et al., 2024; Kim et al., 2024; Huo et al., 2025). For audio-visual LLMs, AVCD (Jung et al., 2025) constructs modality-dependent contrasts, MAD (Chung et al., 2026) adaptively combines modality-conditioned branches, and ASD (Jung et al., 2026) promotes cross-modal information flow. Information-theoretic decoding methods emphasize input-dependent information through context-aware, mutual-information, or PMI-based objectives (Shi et al., 2024; Favero et al., 2024; Chen et al., 2026). Among these methods, MAD (Chung et al., 2026) is most closely related to **RED** because both account for question-dependent evidence requirements. While MAD adaptively weights modality-conditioned branches, **RED** selects among contributions defined relative to a common question-only reference. It decomposes predictive changes into audio, video, and non-additive residual contributions, and adds only the selected contribution to the original audio-visual prediction.

3 Method

3.1 Preliminaries

We consider autoregressive AV-LLMs, where each prediction is conditioned on paired audio-video inputs and a textual query. Given audio a, video v, textual query x, and previously generated tokens $y_{<t}$, let $c_t = (x, y_{<t})$ denote the textual context at decoding step t. For a candidate token $w \in \mathcal{V}$, where $\mathcal{V}$ is the vocabulary, an AV-LLM parameterized by θ assigns the next-token probability $p_\theta(w \mid a, v, c_t)$. Throughout this work, the model parameters θ remain frozen, and all modifications are performed only at decoding time.

Contrastive decoding (CD) adjusts the original prediction using differences between predictions obtained under alternative input conditions (Leng et al., 2024). For example, when the visual input is perturbed from v to v', a representative contrastive score s_{VCD} is

$$s_{\mathrm{VCD}} = (1+\alpha)\log p_\theta(w \mid a, v, c_t) - \alpha \log p_\theta(w \mid a, v', c_t), \tag{1}$$

where α controls the contrastive strength. CD therefore specifies how predictions under two conditions are contrasted, while what the resulting difference represents depends on how the reference condition is constructed.

Among such difference-based formulations, pointwise mutual information (PMI) (Church & Hanks, 1990) is particularly useful when the goal is to quantify the predictive contribution of additional context. PMI measures how much observing z changes the probability assigned to w relative to its marginal probability (*i.e.*, $\mathrm{PMI}(w; z) = \log p(w \mid z) - \log p(w)$). Conditional PMI extends this comparison by holding a shared context fixed. For brevity, we sometimes refer to conditional PMI given c_t simply as PMI when unambiguous. In decoding, this corresponds to comparing next-token predictions with and without an additional source of context while keeping the base context fixed. Following prior works (Fang et al., 2025; Shi et al., 2024), we use the prediction obtained with the perceptual inputs omitted as a practical approximation to the corresponding marginal distribution.

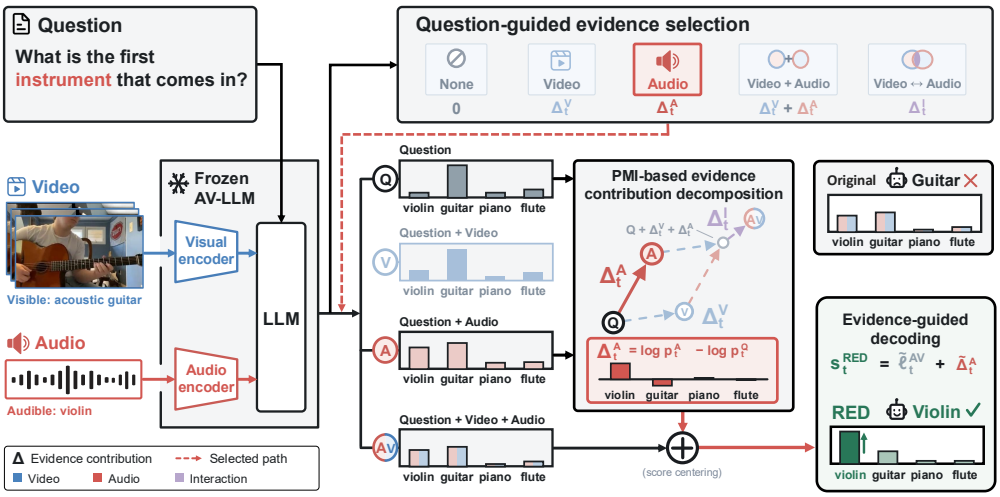


Figure 2: **Overview of RED for an audio-required case.** (i) Given a query, **RED** first identifies the evidence required to answer it, selecting among none, audio, video, audio + video, or audio-visual interaction. (ii) The frozen AV-LLM produces the audio-visual prediction and only the auxiliary predictions required by the selected evidence type, from which RED measures audio, video, and non-additive audio-visual contributions relative to the textual reference. (iii) The selected contribution augments the original audio-visual prediction to produce the final evidence-guided prediction.

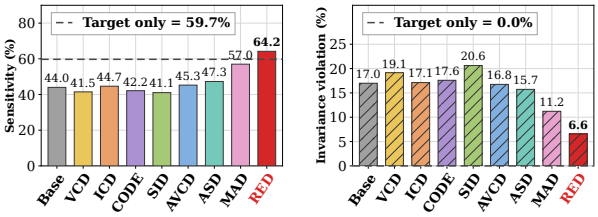
Accordingly, treating the audio-visual input (a, v) as the additional context, we obtain the following conditional PMI approximation:

$$\mathrm{PMI}(w; a, v \mid c_t) \approx \log p_\theta(w \mid a, v, c_t) - \log p_\theta(w \mid c_t). \tag{2}$$

The second term discounts predictive contribution already present in the question and generated context, so the difference captures the additional predictive contribution associated with the perceptual input. However, this joint formulation treats audio and video as a single source of additional context, leaving modality-specific and interaction contributions entangled.

3.2 Motivation

Entangled modality-specific and interaction effects raise a practical question: *does a contrastive correction actually make the prediction follow the evidence required by the question?* Improved accuracy does not necessarily indicate stronger reliance on the required evidence; it may instead arise because non-required perceptual evidence moves the prediction toward the correct answer.



(a) Sensitivity ↑ (b) Invariance violation ↓

Figure 3: **Evidence following under modality interventions.** Higher sensitivity and lower invariance violation indicate better evidence following. The dashed reference uses only the modality required by the question; its invariance violation is zero by construction because the intervened non-required modality is not used. Base receives both audio and video. Results are averaged over question-required evidence types and three backbones.

We therefore use controlled modality interventions to test whether predictions respond to required evidence while remaining stable to non-required evidence. We construct original–contrastive pairs from AVHBench by replacing one modality while keeping the question and the other modality fixed. Replacing the required modality changes the correct answer and defines *sensitivity*, whereas replacing a non-required modality preserves the answer and defines *invariance violation*[1]. Higher sensitivity and lower invari-

[1]Let $y, \hat{y}$ and $y', \hat{y}'$ denote the ground-truth answers and model predictions before and after intervention. Sensitivity counts $\mathbf{1}[\hat{y} = y \wedge \hat{y}' = y']$, while invariance violation counts $\mathbf{1}[\hat{y} \neq \hat{y}']$.

ance violation indicate better evidence following. Details of pair construction, replacement-sample verification, and metric computation are provided in Appendix B.4.

We compare inference with both audio and video against inference using only the modality required by the question. As shown in Figure 3, sensitivity drops from 59.7% when only the required modality is provided to 44.0% when both audio and video are provided, while non-required-modality interventions change 17.0% of predictions. Existing contrastive decoding methods show a similar pattern. Although they can improve final prediction performance, their predictions may still respond weakly to changes in question-required evidence. These results suggest that the required evidence is often available to the model but may not contribute sufficiently when multiple modalities are jointly available. These observations motivate a correction that directly targets the predictive influence of question-relevant evidence, rather than only the final prediction. To capture this influence, we use pointwise mutual information to measure how each evidence source changes the prediction relative to a common reference. Building on these contributions, the question determines which contribution to selectively augment while retaining the original audio-visual prediction.

3.3 Relevant Evidence Decoding

We introduce Relevant Evidence Decoding (**RED**), a training-free method that selectively augments predictive contribution from question-required evidence. Figure 2 provides an overview of **RED**.

PMI-based evidence contribution decomposition. A high modality-conditioned probability need not imply strong support from that modality: the candidate may already be favored by the question and language context (Favero et al., 2024; Wan et al., 2025). We therefore measure each contribution as a change from the same question-only prediction.

For a candidate token w, let $p_t^S(w) = p_\theta(w \mid z_S, x, y_{<t})$ denote the next-token probability under input condition $S \in \{AV, A, V, Q\}$, where $z_{AV} = (a, v)$, $z_A = a$, $z_V = v$, and $z_Q = \varnothing$. All conditions use the same question and generated prefix. These conditions define the possible evidence contributions; which of them is evaluated for a given query is determined by the evidence selection described below.

We define the evidence contribution for S as

$$\Delta_t^S(w) = \log p_t^S(w) - \log p_t^Q(w). \tag{3}$$

Thus, Δ_t^A and Δ_t^V measure the prediction changes induced by audio and video separately, while Δ_t^{AV} measures their joint contribution. We define the remaining difference as the residual audio-visual contribution Δ_t^I:

$$\begin{aligned} \Delta_t^I(w) &= \Delta_t^{AV}(w) - \Delta_t^A(w) - \Delta_t^V(w) \\ &= \log p_t^{AV}(w) + \log p_t^Q(w) - \log p_t^A(w) - \log p_t^V(w). \end{aligned} \tag{4}$$

Equivalently, $\Delta_t^{AV} = \Delta_t^A + \Delta_t^V + \Delta_t^I$, where Δ_t^I denotes the interaction contribution, defined as the non-additive residual under joint conditioning that is not explained by the separate audio and video contributions.

Question-guided evidence selection. **RED** performs a question-only forward pass to determine what perceptual evidence is required to answer the query. We organize the evidence requirements hierarchically by the number of perceptual modalities needed: none, one, or two. With no perceptual input required, the evidence type is $\varnothing$. With one modality required, it is audio (A) or video (V). With two modalities required, we distinguish $A + V$ from $A \leftrightarrow V$. Using only the textual query x, the same frozen backbone selects one of five evidence types:

$$\mathcal{E}(x) \in \{\varnothing,\ A,\ V,\ A + V,\ A \leftrightarrow V\}. \tag{5}$$

For questions requiring both modalities, $A + V$ denotes cases in which information from both audio and video is necessary without explicitly querying the relation between them. In contrast, $A \leftrightarrow V$ denotes cases that explicitly query an audio-video relation, such as correspondence or temporal alignment. The selector is executed once per query before autoregressive decoding, and the selected evidence type is kept fixed across all decoding steps.

The selected evidence type directly determines which predictive contribution is added to the audio-visual prediction. Let $e = \mathcal{E}(x)$ denote the selected evidence type. To avoid tuning hyperparameters on labeled data, we center the scores using unlabeled samples, following prior label-free methods (Zheng et al., 2024; Zhou et al., 2024). We denote the resulting centered audio-visual score and evidence contributions by $\widetilde{\ell}_t^{AV}(w)$ and $\widetilde{\Delta}_t^{S}(w)$, respectively.

RED then selects the corresponding contribution as

$$\rho_t(w) = \begin{cases} 0, & e = \varnothing, \\ \widetilde{\Delta}_t^{e}(w), & e \in \{A, V\}, \\ \widetilde{\Delta}_t^{A}(w) + \widetilde{\Delta}_t^{V}(w), & e = A + V, \\ \widetilde{\Delta}_t^{I}(w), & e = A \leftrightarrow V. \end{cases} \tag{6}$$

The final **RED** score is

$$s_t^{\mathrm{RED}}(w) = \widetilde{\ell}_t^{AV}(w) + \rho_t(w). \tag{7}$$

The resulting score replaces the standard audio-visual next-token score during decoding, while all other decoding operations remain unchanged. **RED** retains the AV branch as the decoding anchor and evaluates only the auxiliary Q, A, and V branches required by the selected evidence type.

4 Experiments

4.1 Experimental setup

Backbones. Following prior works (Chung et al., 2026), we employ Qwen2.5-Omni-7B (Xu et al., 2025) and VideoLLaMA2.1-7B-AV (Cheng et al., 2024) as the main backbone models to enable fair comparisons with existing methods. We additionally evaluate the proposed method with Qwen2.5-Omni-3B to further demonstrate its effectiveness across different model scales. Note that we keep all model parameters frozen.

Benchmarks. We evaluate **RED** across three complementary audio-visual hallucination benchmarks. We evaluate on CMM (Leng et al., 2025) and AVHBench (Kim et al., 2025), both established in prior audio-visual hallucination studies (Chung et al., 2026). CMM examines modality dominance and spurious cross-modal correlations, whereas AVHBench evaluates hallucinations caused by inconsistent audio-visual evidence. We extend this evaluation with SVHalluc (Zhang et al., 2026), covering semantic and temporal hallucinations in speech-vision understanding. For generalization, we additionally evaluate on MUSIC-AVQA (Li et al., 2022) and its challenging MUSIC-AVQA-Hard subset (Kim & Seo, 2026). We provide details on each benchmark in Appendix B.3.

Baselines. We compare **RED** with two groups of training-free decoding baselines: (1) audio-visual hallucination methods, including AVCD (Jung et al., 2025), ASD (Jung et al., 2026), and MAD (Chung et al., 2026); and (2) vision-language hallucination methods adapted to audio-visual inputs, including VCD (Leng et al., 2024), ICD (Wang et al., 2024), CODE (Kim et al., 2024), and SID (Huo et al., 2025). [2] Implementation details are provided in Appendix B.2.

4.2 Main results

Results on CMM. Table 1 shows that **RED** consistently improves CMM across all three backbones, with the largest gain on Qwen2.5-Omni-7B from 82.2% to 89.2%. Notably, the improvement is concentrated on the modality-dominance subsets: **RED** achieves absolute gains of 16.5% and 17.3% on the Vis. and Aud. Dominance subsets, respectively, while changing performance on Lang. only marginally. This pattern indicates that **RED** is particularly effective when perceptual modalities compete to determine the prediction.

Results on AVHBench. **RED** exhibits a similar trend on AVHBench, outperforming the strongest competing decoding methods by 2.1–3.0% across all three backbones. The gain is especially consistent on VDA, where **RED** improves over the corresponding base models by 5.6, 7.9, and 12.1%. Together with the CMM results, this concentration of gains suggests that **RED** is most beneficial in settings where the contribution of perceptual evidence is easily disrupted by other modalities.

[2] All methods are re-evaluated under the same deterministic protocol using the settings reported in the original papers or released implementations.

Table 1: **Performance on audio–visual hallucination benchmarks.** Accuracy (%) on CMM, AVHBench, and SVHalluc. Vis., Aud., and Lang. denote the CMM visual-, audio-, and language-dominance subsets. For SVHalluc, Sem. and Temp. average the semantic (GSA, FGSA, CMSB) and temporal (TA, TF, CMTB) tasks. Avg. is the unweighted mean over all tasks in each benchmark.

Method	CMM				AVHBench				SVHalluc		
	Vis.	Aud.	Lang.	**Avg.**	ADV	VDA	AVM	**Avg.**	Sem.	Temp.	**Avg.**
VideoLLaMA2.1-7B-AV											
Base	73.5	84.0	72.5	82.8	80.4	76.5	57.8	71.6	63.0	43.5	53.3
VCD (Leng et al., 2024)	70.8	84.5	73.8	81.5	**80.7**	74.6	56.3	70.6	64.6	43.6	54.1
ICD (Wang et al., 2024)	73.5	84.2	72.5	82.6	**80.7**	77.0	57.8	71.8	63.2	43.4	53.3
CODE (Kim et al., 2024)	72.5	84.0	72.0	82.4	80.5	76.0	57.8	71.4	63.2	43.1	53.2
SID (Huo et al., 2025)	65.8	83.5	73.0	79.0	80.4	70.4	55.6	68.8	62.7	43.0	52.8
AVCD (Jung et al., 2025)	73.5	82.5	70.8	81.5	**80.7**	76.3	56.9	71.3	63.5	43.4	53.5
ASD (Jung et al., 2026)	74.5	85.5	72.0	82.4	80.0	76.5	56.5	71.0	59.9	41.9	50.9
MAD (Chung et al., 2026)	82.0	82.8	**77.8**	84.7	79.9	80.6	58.5	73.0	68.2	43.7	55.9
RED (Ours)	**83.5**	**88.0**	76.2	**85.6**	78.4	**82.1**	**65.2**	**75.3**	**69.3**	**44.8**	**57.1**
Qwen2.5-Omni-3B											
Base	69.2	84.2	76.5	83.3	80.4	73.3	62.9	72.2	73.8	39.5	56.7
VCD (Leng et al., 2024)	73.2	82.0	76.2	83.3	78.6	72.0	69.7	73.4	72.8	41.5	57.2
ICD (Wang et al., 2024)	68.5	84.0	75.8	82.6	80.4	73.5	63.6	72.5	73.9	40.1	57.0
CODE (Kim et al., 2024)	69.5	85.2	77.2	83.5	80.7	70.6	66.5	72.6	72.1	40.6	56.3
SID (Huo et al., 2025)	71.2	78.0	77.0	82.4	80.1	67.8	67.4	71.8	68.1	43.3	55.7
AVCD (Jung et al., 2025)	70.0	85.0	75.5	82.9	79.8	73.5	65.6	72.9	73.6	41.4	57.5
ASD (Jung et al., 2026)	70.0	84.8	78.0	83.9	82.2	73.7	**70.8**	75.5	73.8	40.5	57.1
MAD (Chung et al., 2026)	74.5	**87.0**	78.8	84.6	82.7	80.0	62.4	75.0	75.1	43.9	59.5
RED (Ours)	**75.8**	**87.0**	**81.5**	**86.2**	**83.6**	**81.2**	70.7	**78.5**	**77.0**	**44.1**	**60.5**
Qwen2.5-Omni-7B											
Base	69.5	71.5	82.2	82.2	80.3	73.4	74.9	76.2	76.1	44.9	60.5
VCD (Leng et al., 2024)	69.0	68.5	82.5	81.7	76.5	72.2	77.1	75.3	74.6	45.2	59.9
ICD (Wang et al., 2024)	68.8	71.2	81.8	82.0	80.5	72.9	76.0	76.5	75.9	44.9	60.4
CODE (Kim et al., 2024)	68.0	68.5	82.8	81.4	78.6	71.1	73.9	74.5	76.5	45.1	60.8
SID (Huo et al., 2025)	73.5	65.8	**83.5**	82.0	77.1	75.7	**77.8**	76.9	73.0	**46.2**	59.6
AVCD (Jung et al., 2025)	69.5	72.8	75.2	80.4	79.3	76.4	73.1	76.3	77.8	45.7	61.7
ASD (Jung et al., 2026)	70.0	74.2	82.0	82.7	84.0	78.9	70.1	77.7	75.3	44.9	60.1
MAD (Chung et al., 2026)	77.0	83.2	83.0	86.1	**84.3**	78.8	76.4	79.8	81.4	45.1	63.2
RED (Ours)	**86.0**	**88.8**	83.0	**89.2**	84.0	**85.5**	76.3	**81.9**	**81.9**	44.9	**63.4**

Results on SVHalluc. **RED** achieves the best SVHalluc accuracy across all three backbones, improving standard decoding from 53.3–60.5% to 57.1–63.4% and consistently outperforming competing decoding methods. Unlike prior audio-visual hallucination benchmarks that primarily rely on environmental sounds as indicators of event occurrence, SVHalluc requires grounding semantically rich speech with corresponding visual evidence (Zhang et al., 2026). The consistent advantage of **RED** in this more challenging speech–vision setting suggests that selectively targeting question-required predictive contributions remains effective beyond simpler acoustic-event grounding.

Results on general audio-visual QA. Table 2 evaluates whether these gains extend beyond hallucination-specific benchmarks. **RED** improves Qwen2.5-Omni-7B from 64.5% to 67.1% on MUSIC-AVQA and from 31.9% to 40.5% on MUSIC-AVQA-Hard, matching MAD (Chung et al., 2026) on the former and exceeding it by 4.5% on the latter. The substantially larger gain on MUSIC-AVQA-Hard indicates that the benefit of **RED** persists, and becomes more pronounced, as audio-visual reasoning becomes more challenging.

4.3 Analyses

True- and false-positive analysis. Accuracy on binary hallucination benchmarks can be confounded by response bias, making it difficult to determine whether performance gains arise from improved perceptual grounding or simply a shifted decision tendency (Yin et al., 2025). To disentangle these effects, we separately evaluate the true-positive rate (TPR) and false-positive rate (FPR), as shown in Figure 4.

Table 2: **General audio-visual QA with Qwen2.5-Omni-7B.** Accuracy (%) and relative TTFT (× Base) on MUSIC-AVQA (Full) and MUSIC-AVQA-Hard (Hard).

Method	Full		Hard	
	Acc.	TTFT	Acc.	TTFT
Base	64.5	1.0×	31.9	1.0×
VCD	64.5	2.8×	32.4	2.8×
SID	62.9	1.4×	31.5	1.4×
ICD	65.7	3.1×	32.7	3.0×
CODE	63.7	2.7×	32.3	2.7×
AVCD	64.6	2.0×	31.7	2.0×
ASD	62.4	4.1×	28.2	4.2×
MAD	**67.1**	2.7×	36.0	2.6×
RED (Ours)	**67.1**	1.6×	**40.5**	1.7×

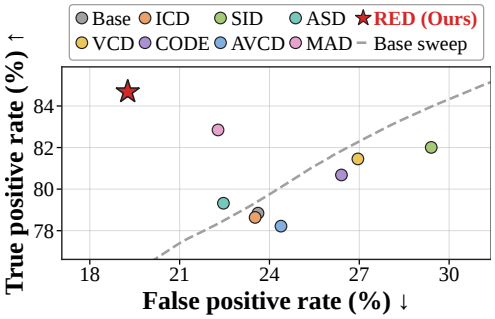


Figure 4: **True-positive and false-positive rates.** **RED** achieves the highest TPR and the lowest FPR among the compared methods. The Base threshold sweep (dashed) does not attain **RED**'s TPR at the same or lower FPR.

RED improves both rates simultaneously, increasing the mean TPR from 78.8% to 84.7% while reducing the mean FPR from 23.6% to 19.3%. Among methods that improve both rates relative to Base, **RED** yields the largest improvement on each axis, reducing FPR by 4.3% while increasing TPR by 5.9%. The dashed Base sweep varies the Base decision threshold on the same items, with the Base marker at threshold zero; no threshold reaches **RED**'s TPR at **RED**'s FPR.

Component ablation. In Table 3, we ablate question-guided evidence selection (*Q-guided*) and PMI-based contribution decomposition (*PMI*). With only *PMI* enabled, we add the audio, video, and residual interaction contributions to every question. With only *Q-guided* enabled, we replace PMI contributions with raw modality-conditioned scores: audio-only for Audio, video-only for Video, and the full audio-visual score for Both and Relation. With neither component enabled, we add the raw audio-only and video-only scores to the audio-visual score. Both components are necessary for the full **RED** formulation: on AVHBench, adding PMI decomposition to Q-guided selection improves accuracy from 76.2% to 78.6% and AUROC from 82.7% to 86.1%. The full model also achieves the strongest overall results, supporting the role of PMI decomposition in isolating the predictive contribution to be selectively reinforced.

Table 3: **Component analysis of evidence correction.** We ablate question-guided evidence selection and PMI-based contribution decomposition with the remaining RED procedure fixed. Results are averaged over three backbones. AUC denotes AUROC.

Q-guided	PMI	CMM		AVH	
		Acc.	AUC	Acc.	AUC
✗	✗	84.8	92.3	75.2	82.0
✓	✗	86.8	93.6	76.2	82.7
✗	✓	84.2	91.9	75.8	82.8
✓	✓	**87.0**	**94.0**	**78.6**	**86.1**

Evidence selection. To examine sensitivity to the selector implementation, Table 4 compares Base, *Selected target-only*, and alternative evidence assignments. *Selected target-only* uses the selected perceptual input, or full audio-visual input for Both and Relation, and substantially improves over Base; **RED** retains the full audio-visual prediction and further improves it on both MUSIC-AVQA benchmarks.

For correction variants, only the evidence assignment changes: *Mismatched* averages the other three non-None types, *Random* averages five uniform assignments, *Heuristic* maps explicit question cues to evidence types, and *Oracle* uses available question-level evidence labels as a label-informed upper-bound reference. Mismatched and Random underperform **RED**, whereas the competitive Heuristic and Oracle results show that correctly aligning the correction with question-required evidence is more important than the particular selector implementation. **RED** performs this selection with the same frozen backbone without an explicit rule-based mapping.

Inference efficiency. Figure 5 compares accuracy and relative TTFT averaged across nine backbone-benchmark settings. **RED** improves Base accuracy from 71.0% to 75.3% at 1.5× TTFT.

Table 4: **Evidence selection analysis on Qwen2.5-Omni-7B.** Accuracy (%) for Base, Selected target-only, and alternative evidence-assignment variants. Full and Hard denote MUSIC-AVQA and MUSIC-AVQA-Hard.

Variant	AVH	CMM	MUSIC-AVQA	
			Full	Hard
Base	76.2	82.2	64.5	31.9
Selected target-only	81.7	88.3	65.4	37.0
Mismatched	75.4	82.1	59.5	36.8
Random	78.0	84.0	62.6	37.7
Heuristic	**81.9**	**89.2**	66.9	40.7
RED	**81.9**	**89.2**	67.1	40.5
Oracle	–	–	**67.7**	**41.6**

"–" denotes unavailable question-level oracle evidence labels.

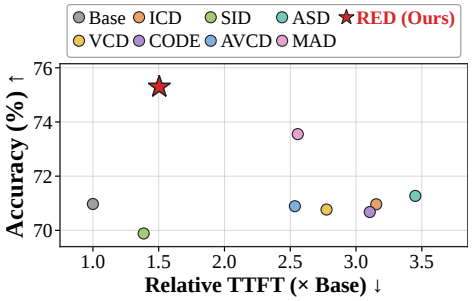


Figure 5: **Accuracy versus relative TTFT.** Means over CMM, AVHBench, and SVHalluc across three backbones. TTFT is normalized to the corresponding Base within each setting.

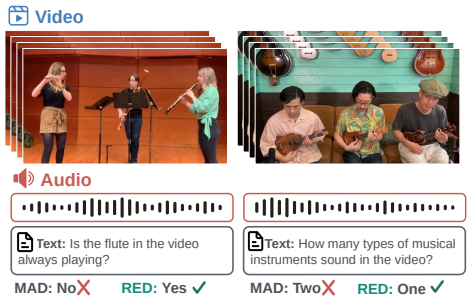
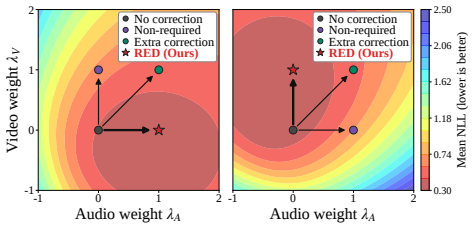


Figure 6: **Qualitative examples on MUSIC-AVQA-Hard.** **RED** correctly answers examples where MAD fails.



(a) Audio-required questions (b) Video-required questions

Figure 7: **Loss landscape on CMM.** Mean binary-answer NLL over audio- and video-required subsets, averaged across three backbones. **RED** corresponds to $(\lambda_A, \lambda_V) = (1, 0)$ and $(0, 1)$, respectively.

Compared with MAD (Chung et al., 2026), **RED** improves accuracy from 73.6% to 75.3% while reducing TTFT from 2.6× to 1.5×, a 1.7× speedup.

Qualitative and contribution analyses. Figure 6 illustrates modality bias in audio-visual decoding and shows that selectively augmenting question-required evidence can correct such failures without additional training. Figure 7 further analyzes this behavior through the loss landscape obtained by scaling the audio and video contributions with λ_A and λ_V, respectively, while keeping evidence selection fixed. In both cases, **RED** yields lower NLL than the three reference settings, supporting correction aligned with question-required evidence. See Appendix A.3 for further details.

5 Conclusion

In this work, we study cross-modal hallucination in AV-LLMs through the lens of *evidence following*, showing that correct predictions can remain weakly dependent on the perceptual evidence required by the question. We introduce Relevant Evidence Decoding (**RED**), a training-free method that decomposes audio-visual predictive changes relative to a common question-only reference and selectively augments the contribution corresponding to question-required evidence. Across multiple AV-LLMs and benchmarks, **RED** consistently improves hallucination mitigation and general audio-visual reasoning, while controlled interventions show increased sensitivity to required evidence and reduced sensitivity to irrelevant evidence. These results suggest that effective multimodal decoding should not only improve prediction accuracy, but also ensure that predictions respond to the evidence that should determine the answer.

AI USE STATEMENT

During manuscript preparation, large language models were used for writing assistance to improve grammar, wording, and readability of text drafted by the authors. We also used feedback from Google's Paper Assistant Tool (PAT), provided through ICLR, to support manuscript revision. The research ideas, methodological design, experimental setup, data analysis, and interpretation of results were carried out by the authors. The authors reviewed the final manuscript and take full responsibility for its content and claims.

ETHICS STATEMENT

Cross-modal hallucinations can reduce the reliability of multimodal systems, particularly in applications such as assistive technologies, multimodal assistants, and audio-visual content understanding. By improving the use of question-relevant evidence, **RED** may help make existing AV-LLMs more dependable in real-world settings. As a training-free method, **RED** requires no task-specific retraining, reducing the computational burden and limiting the need to use potentially sensitive audio-visual data for additional model optimization. These properties may facilitate hallucination mitigation in resource-constrained and privacy-sensitive settings, supporting more reliable use of existing multimodal systems.

REPRODUCIBILITY STATEMENT

Our experiments are conducted on NVIDIA RTX A6000 GPUs. Details required for reproducing the experiments, including dataset settings, evaluation protocols, baseline implementations, hyperparameter configurations, and the implementation details of **RED**, are provided in the main text and Appendix. We plan to release the code and supporting implementation upon publication.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pp. 23716–23736, 2022. 1

Ami Baid, Zihui Xue, and Kristen Grauman. Don't let the video speak: Audio-contrastive preference optimization for audio-visual language models. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2026. 1, 3

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. 1

Ashutosh Chaubey, Jiacheng Pang, and Mohammad Soleymani. Mod-dpo: Towards mitigating cross-modal hallucinations in omni llms using modality decoupled preference optimization. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pp. 18284–18294, 2026. 3

Dongsheng Chen, Yingqi Zhu, Xingyue Zhang, Wenqing Zhou, and Lei Li. Dynamic PMI-guided contrastive decoding reduces hallucination in large language models: A unified framework of fine-grained input transformations. In *Findings of the Association for Computational Linguistics (ACL Findings)*, pp. 24223–24235, 2026. 3, 26

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 1, 2, 6

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 *Proc. of Int'l Conf. on Learning Representations (ICLR)*, 2024. 1

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 *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pp. 40175–40185, 2026. 1, 2, 3, 6, 7, 9, 14, 20, 21, 23, 25, 26, 27, 28

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990. 2, 3

Hao Fang, Changle Zhou, Jiawei Kong, Kuofeng Gao, Bin Chen, Tao Liang, Guojun Ma, and Shu-Tao Xia. Grounding language with vision: A conditional mutual information calibrated decoding strategy for reducing hallucinations in lvlms. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2025. 3

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14312, 2024. 1, 2, 3, 5, 26

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics (TACL)*, 10:1138–1158, 2022. 22

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 1307–1323, 2020. 22

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, 2024. 3

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *Proc. of Int'l Conf. on Learning Representations (ICLR)*, 2025. 3, 6, 7, 19, 20, 21, 25, 26, 27, 28

Chaeyoung Jung, Youngjoon Jang, and Joon Son Chung. Avcd: Mitigating hallucinations in audio-visual large language models through contrastive decoding. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2025. 1, 2, 3, 6, 7, 14, 20, 21, 25, 26, 27, 28

Jihoo Jung, Chaeyoung Jung, Ji-Hoon Kim, and Joon Son Chung. Probing cross-modal information hubs in audio-visual llms. In *Proc. of Int'l Conf. on Machine Learning (ICML)*, 2026. 1, 3, 6, 7, 20, 21, 25, 26, 27, 28

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *Proc. of Int'l Conf. on Learning Representations (ICLR)*, 2020. 22

Geewook Kim and Minjoon Seo. Do modern video-llms need to listen? a benchmark audit and scalable remedy. In *Proc. of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2026. 2, 6, 22

Junho Kim, Hyun Jun Kim, Yeon Ju Kim, and Yong Man Ro. Code: Contrasting self-generated description to combat hallucination in large multi-modal models. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2024. 1, 3, 6, 7, 20, 21, 25, 26, 27, 28

Sung-Bin Kim, Hyun-Bin Oh, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. In *Proc. of Int'l Conf. on Learning Representations (ICLR)*, 2025. 2, 3, 6, 21, 22

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Trans. on Audio, Speech, and Language Processing (TASLP)*, 28:2880–2894, 2020. 23

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 *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, 2024. 1, 3, 6, 7, 20, 21, 25, 26, 27, 28

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. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2025. 1, 2, 3, 6, 21

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pp. 13723–13733, 2025. 1

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pp. 19108–19118, 2022. 2, 6, 22

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. 1

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proc. of Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 292–305, 2023. 3

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pp. 34892–34916, 2023. 1

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2023. 23

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4902–4912, 2020. 22

Xuyang Shen, Dong Li, Jinxing Zhou, Zhen Qin, Bowen He, Xiaodong Han, Aixuan Li, Yuchao Dai, Lingpeng Kong, Meng Wang, Yu Qiao, and Yiran Zhong. Fine-grained audible video description. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pp. 10585–10596, 2023. 27

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proc. of the Conf. of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 783–791, 2024. 1, 2, 3, 26

Zifu Wan, Ce Zhang, Silong Yong, Martin Q. Ma, Simon Stepputtis, Louis-Philippe Morency, Deva Ramanan, Katia Sycara, and Yaqi Xie. Only: One-layer intervention sufficiently mitigates hallucinations in large vision-language models. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, pp. 3225–3234, 2025. 5

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 Findings)*, pp. 15840–15853, 2024. 1, 3, 6, 7, 20, 21, 25, 26, 27, 28

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pp. 13204–13214, 2024. 1, 22

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. *arXiv preprint arXiv:2503.20215*, 2025. 1, 2, 6

Qilang Ye, Wei Zeng, Meng Liu, Jie Zhang, Yupeng Hu, Zitong Yu, and Yu Zhou. When eyes and ears disagree: Can mllms discern audio-visual confusion? In *Proc. of the AAAI Conf. on Artificial Intelligence*, pp. 11955–11963, 2026. 1

Hao Yin, Guangzong Si, and Zilei Wang. The mirage of performance gains: Why contrastive decoding fails to mitigate object hallucinations in mllms? In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2025. 1, 3, 7, 22

Chenshuang Zhang, Kyeong Seon Kim, Chengxin Liu, and Tae-Hyun Oh. Svhalluc: Benchmarking speech-vision hallucination in audio-visual large language models. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2026. 2, 3, 6, 7, 21, 24

Chen Zhao, Jiajun Ma, Qilong Huang, Tiehan Fan, Hongyu Li, Zhuoliang Kang, Xiaoming Wei, Jian Yang, and Ying Tai. Temporal and cross-modal alignment for enhanced audiovisual video captioning. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2026a. 27

Feiyu Zhao, Yiming Chen, Wenhuan Lu, Daipeng Zhang, Xianghu Yue, and Jianguo Wei. Halluaudio: A comprehensive benchmark for hallucination detection in large audio-language models. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 38797–38816, 2026b. 1

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *Proc. of Int'l Conf. on Learning Representations (ICLR)*, 2024. 6, 14

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine Heller, and Subhrajit Roy. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. In *Proc. of Int'l Conf. on Learning Representations (ICLR)*, 2024. 6, 14

Appendix

This appendix provides additional method details, experimental protocols, supplementary results, and qualitative analyses. The contents are organized as follows:

- Appendix A: Method details and inference analysis;
- Appendix B: Experimental setup, baselines, benchmarks, and intervention details;
- Appendix C: Additional quantitative results;
- Appendix D: Qualitative analyses;
- Appendix E: Limitations.

A Method details

A.1 Question-guided evidence selection

RED determines the evidence required by the question using the same frozen backbone and the question text alone. As shown in Figure A, the model assigns each question to one of five evidence types—None, Audio, Video, Both, and Relation—corresponding to $\varnothing$, A, V, $A+V$, and $A \leftrightarrow V$ in the main text. No audio or video input is provided during this selection step. The selected evidence type determines which contribution is subsequently added to the audio-visual prediction.

A.2 Effect of score centering

To match the sample budget used in prior decoding protocols, which select decoding strengths using 100 labeled examples per dataset (Jung et al., 2025; Chung et al., 2026), we use the same number of samples for score centering without labels. Following prior label-free methods (Zheng et al., 2024; Zhou et al., 2024), we estimate fixed mean offsets from these unlabeled questions. Let $\ell_t^{AV}(w) = \log p_t^{AV}(w)$ denote the uncentered audio-visual log-probability score. The centered scores are $\widetilde{\ell}_t^{AV}(w) = \ell_t^{AV}(w) - \eta_e^{AV}(w)$ and $\widetilde{\Delta}_t^{S}(w) = \Delta_t^{S}(w) - \eta_e^{S}(w)$, where $S \in \{A, V, I\}$. Here, e denotes the selected evidence type, and $\eta_e^{AV}(w)$ and $\eta_e^{S}(w)$ are the means of the corresponding audio-visual score and evidence contribution for candidate w, computed over the sampled questions assigned to e for each backbone and benchmark. When fewer than five valid samples are available, we use the mean of the corresponding quantity pooled across evidence types. The offsets are kept fixed throughout evaluation.

Table A isolates the effect of mean-offset centering on CMM while fixing the evidence assignments, contribution definitions, and evaluation population. Centering improves accuracy from 87.3% to 89.2% on Qwen2.5-Omni-7B, while producing smaller changes on Qwen2.5-Omni-3B and VideoLLaMA2.1-7B-AV. In contrast, AUROC changes only marginally across all three backbones, from 95.4% to 95.5%, 93.6% to 93.6%, and 92.5% to 92.8%, respectively. This pattern suggests that centering primarily corrects backbone-dependent systematic score offsets at the final decision stage, rather than substantially changing the relative discrimination induced by the evidence-based scores.

Sensitivity to sample size. Figure B examines the effect of using 25, 50, 100, 200, or 400 unlabeled questions to estimate the centering offsets on CMM across three backbones. We keep the sampling seed, evidence assignments, and backbone predictions fixed, varying only the sample budget. Increasing the budget does not improve performance monotonically: Qwen2.5-Omni-3B benefits from larger budgets, while Qwen2.5-Omni-7B and VideoLLaMA2.1-7B-AV show smaller accuracy changes beyond 100 questions.

Robustness to sample selection. To assess variability at the default budget, we independently sample 100 unlabeled questions using five random seeds and recompute the offsets for each draw, keeping all other settings fixed. Following our evaluation protocol, each draw is evaluated on all 2,400 CMM questions, including the sampled questions. Figure C reports the individual results and their mean and standard deviation alongside the reported **RED** results. Accuracy is $86.20 \pm 0.51\%$, $89.54 \pm 0.31\%$, and $85.08 \pm 0.85\%$ for Qwen2.5-Omni-3B, Qwen2.5-Omni-7B, and VideoLLaMA2.1-7B-AV, respectively. The accuracy improvement from centering persists across all

Task
Classify the following question about a video that has an audio track into **exactly one category** from: **[Audio, Video, Both, Relation, None]**, based on which evidence is needed to answer it.

Consider:
Audio: the question is decided by **listening alone** — whether a sound occurs, is audible, or a named thing makes sound, even if that thing is also visible.
Video: the question is decided by **watching alone** — whether something is visible, appears, or happens on screen, even if that thing also makes sound.
Both: the question needs listening and watching together — it attributes sounds to things that must be **identified by sight**, such as which one is louder or sounds first, how many are actually sounding, or where the sounding one is; or it asks **two separable facts**, one heard and one seen.
Relation: the question **explicitly compares the audio content with the visual content** — whether they match, correspond, or describe the same event.
None: the question needs **neither listening nor watching**.

Examples:
"Is the espresso machine making sound in the audio?" → **Audio** (only whether the named thing makes any sound, even if it is visible).
"Does a bicycle appear anywhere on screen?" → **Video**.
"Is the harmonica louder than the organ?" → **Audio** (named targets, hearing alone suffices).
"Is the performer on the left louder than the performer on the right?" → **Both** (targets identified by sight).
"How many performers appear on screen?" → **Video**.
"How many of the performers on screen are actually producing sound?" → **Both**.
"Do the sounds and the visual scene describe the same event?" → **Relation**.
"Do the audio and the video match each other?" → **Relation**.
"Did you see shape of the snowman is round in the video?" → **Video** (a visual attribute, decided by watching alone).

Question
Classify the following question: "**{q}**"
Provide **only the category**.

Figure A: **Question-guided evidence selection.** The frozen backbone identifies the required evidence using the question text alone.

Table A: **Effect of score centering on CMM.** Task-macro accuracy and AUROC (%) for **RED** with and without centering.

Method	Centering	Acc.	AUROC
VideoLLaMA2.1-7B-AV			
RED (Ours)	×	85.5	92.5
RED (Ours)	✓	**85.6**	**92.8**
Qwen2.5-Omni-3B			
RED (Ours)	×	86.1	**93.6**
RED (Ours)	✓	**86.2**	**93.6**
Qwen2.5-Omni-7B			
RED (Ours)	×	87.3	95.4
RED (Ours)	✓	**89.2**	**95.5**

five draws for Qwen2.5-Omni-7B, whereas the other backbones show less consistent benefits, with greater sampling variability for VideoLLaMA2.1-7B-AV.

A.3 Diagnostics of evidence contribution selection

RED associates each selected evidence type with a corresponding centered predictive contribution: $A \to \widetilde{\Delta}^A$, $V \to \widetilde{\Delta}^V$, $A + V \to \widetilde{\Delta}^A + \widetilde{\Delta}^V$, and $A \leftrightarrow V \to \widetilde{\Delta}^I$. For brevity, we omit the tildes in the figures and tables of this subsection; all reported correction terms use the centered contributions

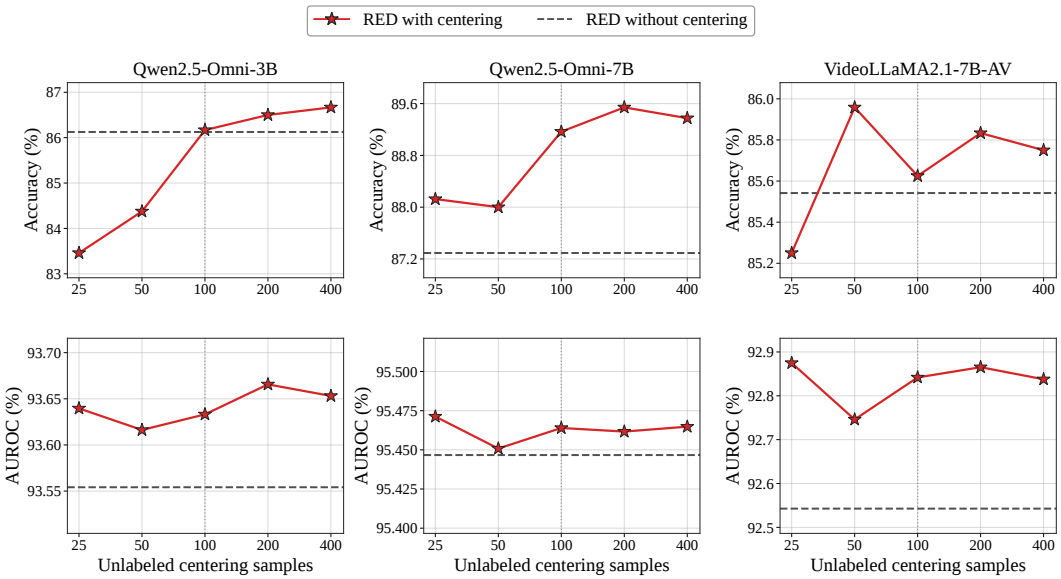
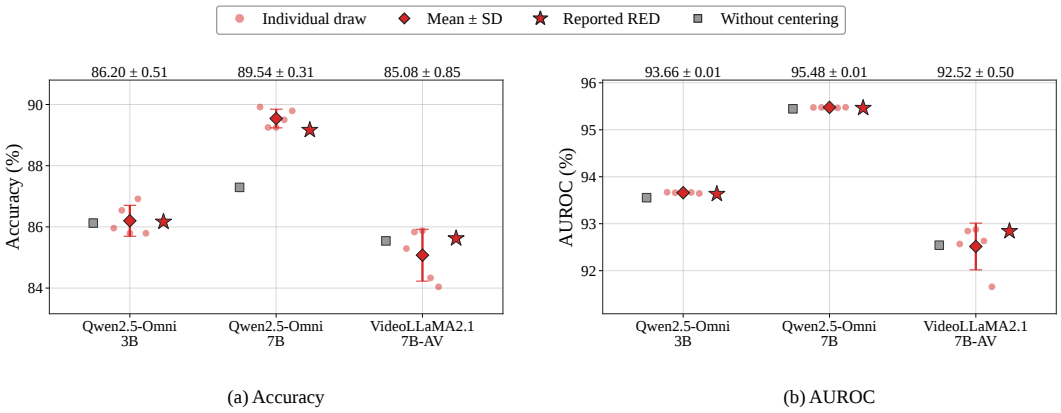


Figure B: **Sensitivity to centering sample size on CMM.** We vary the number of unlabeled centering questions under a fixed sampling configuration. Dashed lines denote **RED** without centering; the dotted vertical line marks the default budget of 100 questions.



(a) Accuracy (b) AUROC

Figure C: **Robustness to centering sample selection on CMM.** Small circles show results from five random draws of 100 unlabeled questions. Diamonds and error bars indicate the mean and one sample standard deviation. Stars show the reported **RED** results, and squares denote **RED** without centering.

defined in the main text. We examine whether prediction quality depends on which contribution is added to the audio-visual score. Within each comparison, we keep the question-guided evidence selection, auxiliary predictions, and centering offsets fixed, and vary only the added contribution. This isolates the effect of contribution choice from changes in evidence selection.

Contribution choice matters. Figure D examines how the choice of evidence contribution affects the true-positive and true-negative operating point. For all correction-based variants, we keep the question-guided evidence selection, auxiliary predictions, and centering procedure fixed and vary only the contribution added to the audio-visual score. *Base* denotes standard decoding without score centering or an additional evidence contribution. *No correction* uses the centered audio-visual score without an additional evidence contribution, whereas **RED** adds the contribution corresponding to the selected evidence type. *Mismatched correction* keeps the selected evidence type fixed but adds

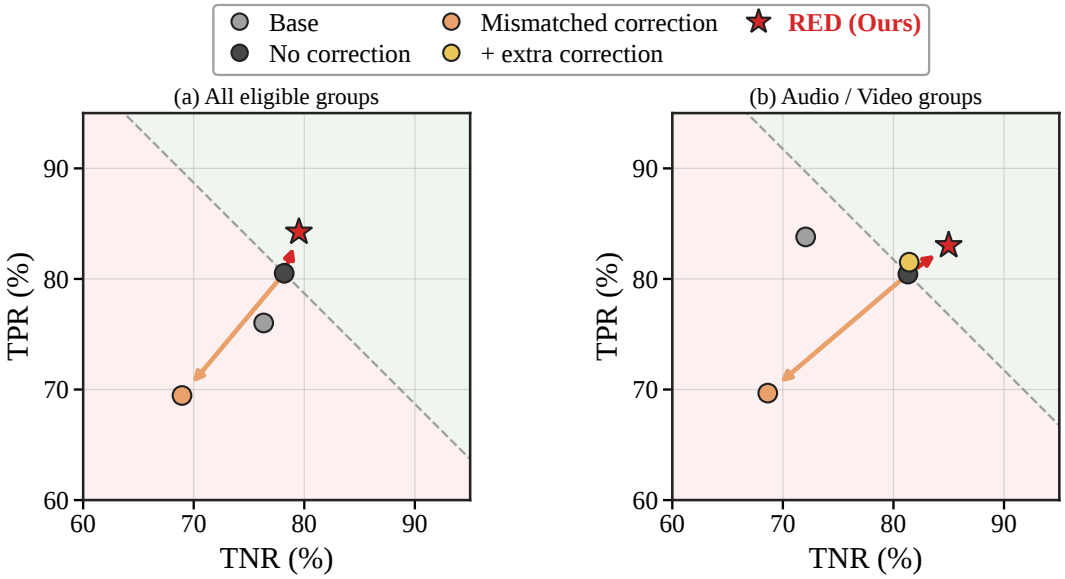


Figure D: **Effect of choosing different evidence contributions.** We keep the evidence selection and centering fixed and vary only the contribution added to the audio-visual score. The left panel includes all eligible groups, while the right panel focuses on Audio and Video groups and additionally compares adding the other unimodal contribution. *No correction* denotes the centered audio-visual score without an additional evidence contribution. Higher TPR and TNR are better.

a non-corresponding contribution; this differs from the mismatched evidence-assignment strategy analyzed in the main text.

Across the eligible groups, **RED** moves the operating point toward higher TPR and TNR, whereas mismatched corrections move it in the opposite direction. For Audio and Video groups, we additionally consider + *extra correction*, which adds both unimodal contributions rather than only the selected one. Its limited benefit compared with **RED** indicates that adding more modality contributions is not necessarily beneficial; the correction should match the evidence required by the question.

Loss landscape on CMM. Figure 7 shows binary-answer negative log-likelihood (NLL) as the audio and video contributions are scaled. For each question, we form

$$s_{\lambda_A,\lambda_V}(w) = \widetilde{\ell}^{AV}(w) + \lambda_A \widetilde{\Delta}^{A}(w) + \lambda_V \widetilde{\Delta}^{V}(w), \qquad w \in \{\text{Yes}, \text{No}\},$$

and compute the NLL of the reference answer after normalizing these two scores with a softmax. We vary λ_A and λ_V over $[-1, 2]$ with a step of 0.025, while keeping the evidence selections, underlying predictions, and centering offsets fixed. The offsets are estimated from 100 unlabeled CMM questions using sampling seed 5, following Appendix A.2.

The analysis uses questions with all four prediction branches available from the VAL, Vis., and Aud. subsets of CMM. All four branches are required because the two-dimensional sweep evaluates both $\widetilde{\Delta}^{A}$ and $\widetilde{\Delta}^{V}$ for the same question, including the non-selected contribution that **RED** would not normally evaluate. We group these questions by each backbone's question-only evidence selection, retaining the Audio and Video groups. For Qwen2.5-Omni-3B, Qwen2.5-Omni-7B, and VideoLLaMA2.1-7B-AV, respectively, the Audio groups contain 618, 600, and 513 questions, and the Video groups contain 582, 600, and 686 questions. At each coefficient pair, we first average NLL over questions within each backbone's group and then average equally across the three backbones.

The no-correction setting is $(\lambda_A, \lambda_V) = (0, 0)$ and uses only the centered audio-visual score. **RED** corresponds to $(1, 0)$ for Audio and $(0, 1)$ for Video; non-required correction uses the opposite modality contribution, and extra correction corresponds to $(1, 1)$. Note that **RED** uses the fixed coefficients implied by its decoding rule; the landscape is a diagnostic analysis and is not used to tune or select decoding strengths.

Table B: **Modality-specific and residual contributions on AVHBench.** Accuracy and AUROC (%) with Qwen2.5-Omni-3B when adding the audio (Δ^A), video (Δ^V), or residual audio-visual (Δ^I) contribution to the same centered audio-visual score. VDA, ADV, and AVM correspond to Audio (A), Video (V), and Relation ($A \leftrightarrow V$) evidence requirements, respectively. The centering offsets are fixed across variants. **Bold** indicates the best result in each row for each metric.

Task Evidence	$+\Delta^A$		$+\Delta^V$		$+\Delta^I$	
	Acc.	**AUROC**	**Acc.**	**AUROC**	**Acc.**	**AUROC**
VDA A	**81.27**	**89.24**	70.39	76.89	46.38	44.60
ADV V	77.99	84.26	**83.63**	**88.82**	64.08	68.20
AVM $A \leftrightarrow V$	69.56	77.65	62.42	68.51	**70.68**	**78.54**

Table C: **Additive contributions for both-modality evidence on MUSIC-AVQA-Hard.** Accuracy (%) with Qwen2.5-Omni-7B. Prompt-selected contains questions assigned to $A + V$ by **RED**'s question-guided evidence selection, whereas Oracle-tagged contains questions identified by the dataset's Audio-Visual evidence tags. All variants use the same centering offsets and differ only in the contribution added to the centered audio-visual score. **Bold** indicates the best result in each row.

$A + V$ **subset**	**n**	$+\Delta^A$	$+\Delta^V$	$+(\Delta^A + \Delta^V)$	$+\Delta^I$
Prompt-selected	1,786	36.73	**38.63**	38.19	33.82
Oracle-tagged	2,124	37.95	38.79	**40.16**	30.23

Modality-specific and residual contributions. Table B examines Audio, Video, and Relation evidence requirements on AVHBench. For VDA, which requires audio evidence, Δ^A achieves the highest accuracy and AUROC. For ADV, which requires video evidence, Δ^V performs best on both metrics. For AVM, which evaluates the relation between audio and visual content, the residual audio-visual contribution Δ^I achieves the highest accuracy and AUROC. The preferred contribution therefore matches the corresponding A, V, and $A \leftrightarrow V$ evidence requirements.

Additive contributions for both-modality evidence. The $A + V$ evidence type is too sparsely represented on AVHBench for a meaningful direct analysis. We therefore examine this case on MUSIC-AVQA-Hard, which also provides dataset evidence annotations identifying questions that require both audio and video.

Table C compares questions selected as $A + V$ by **RED**'s question-guided evidence selection with questions identified by the dataset's Audio-Visual evidence tags. On the Oracle-tagged subset, $\Delta^A + \Delta^V$ achieves the highest accuracy at 40.16%, outperforming either modality-specific contribution alone. This result supports the additive contribution when the evidence requirement explicitly involves both modalities.

The Prompt-selected subset exhibits a different pattern: Δ^V slightly outperforms $\Delta^A + \Delta^V$ (38.63% vs. 38.19%). Because **RED** determines the required evidence from the textual query alone, the questions it assigns to $A + V$ need not coincide with those identified by the dataset annotations. Accordingly, the effectiveness of selective decoding depends not only on mapping an evidence type to its predictive contribution, but also on identifying the evidence required by the question.

Evidence selection across benchmarks. Figure E summarizes question-only evidence selections across five benchmarks, averaged equally over Qwen2.5-Omni-3B, Qwen2.5-Omni-7B, and VideoLLaMA2.1-7B-AV. Each backbone's distribution includes all questions in the corresponding evaluation split: 2,400 for CMM, 5,302 for AVHBench, 2,405 for SVHalluc, 9,129 for the officially corrected MUSIC-AVQA test split, and 2,508 for MUSIC-AVQA-Hard. CMM predominantly selects Audio or Video, while Relation accounts for a substantial share on AVHBench. SVHalluc shows a broader distribution over Video, Both, and Relation. Both is the most frequently selected type on MUSIC-AVQA and MUSIC-AVQA-Hard, especially on the latter. These distributions describe the selectors' predicted evidence requirements rather than ground-truth modality dependencies.

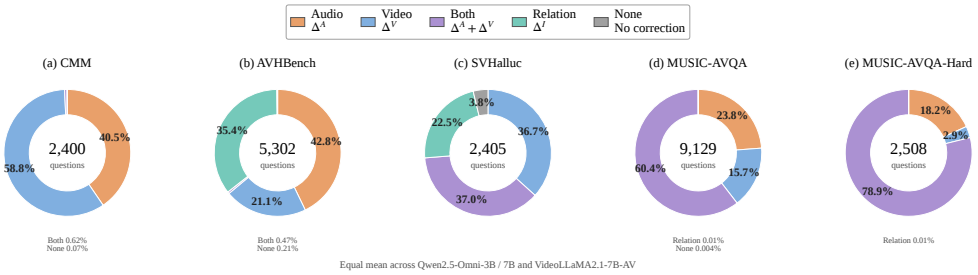


Figure E: **Question-guided evidence selection across five benchmarks.** For each backbone, proportions are computed over all questions in each evaluation split and then averaged equally across three backbones. Centers report the number of questions per backbone. Audio, Video, Both, Relation, and None select Δ^A, Δ^V, $\Delta^A + \Delta^V$, Δ^I, and no additional correction, respectively. Both selects additive audio and video contributions; Relation selects their residual interaction. The distributions reflect question-only selector predictions, not oracle annotations. Nonzero categories below 1% are listed beneath each chart. MUSIC-AVQA uses the official corrected test annotations (9,129 questions); MUSIC-AVQA-Hard contains 2,508 questions from this split.

Table D: **Runtime breakdown of RED.** Stage times are averaged equally over CMM, AVHBench, and SVHalluc across three backbones using separate profiling runs.

Stage	Time (ms)
Evidence selection	89.1
Media preparation	2372.1
Perceptual encoding	1881.4
Answer-prompt prefill	1381.7
Answer-token prediction	28.4
Score composition and overhead	37.2
RED (Ours)	**5789.9**

A.4 Inference Efficiency Analysis

Inference stages. Table D profiles **RED** across three backbones on CMM, AVHBench, and SVHalluc. The selector predicts the evidence type once per question, while the required answer-prediction paths are evaluated separately. Evidence selection incurs little overhead, whereas media preparation and perceptual encoding account for most of the runtime. By evaluating only the auxiliary prediction paths required by the selected evidence type, **RED** avoids unnecessary perceptual processing.

TTFT across benchmarks. Table E reports relative TTFT for each backbone and hallucination benchmark. **RED** requires approximately $1.2\times$–$1.7\times$ Base TTFT and remains faster than MAD and ASD across all nine settings, consistent with evaluating only the auxiliary prediction paths required by the selected evidence type.

Accuracy–TTFT trade-off under SID pruning. Because SID (Huo et al., 2025) can reduce inference cost by pruning perceptual tokens, we compare its accuracy–TTFT trade-off under a common 2-fps configuration. As shown in Table F, more aggressive pruning reduces SID's TTFT but also lowers its average CMM accuracy. With 90% of audio and visual tokens removed, SID reaches 81.54% accuracy at $1.509\times$ Base TTFT, whereas **RED** reaches 85.00% at $1.433\times$ under the same configuration.

Table E: **Relative TTFT across audio-visual hallucination benchmarks.** Time to first token normalized to the corresponding Base (× Base; lower is better).

Method	VideoLLaMA2.1-7B-AV			Qwen2.5-Omni-3B			Qwen2.5-Omni-7B		
	CMM	AVH	SVH	CMM	AVH	SVH	CMM	AVH	SVH
Base	1.0×	1.0×	1.0×	1.0×	1.0×	1.0×	1.0×	1.0×	1.0×
VCD (Leng et al., 2024)	2.2×	3.0×	1.6×	2.9×	3.2×	2.9×	3.0×	3.3×	3.0×
ICD (Wang et al., 2024)	2.1×	2.8×	1.6×	3.8×	4.1×	3.0×	3.7×	4.1×	3.2×
CODE (Kim et al., 2024)	2.7×	3.8×	2.0×	3.6×	3.8×	2.1×	4.0×	3.8×	2.2×
SID (Huo et al., 2025)	1.2×	1.4×	1.1×	1.5×	1.5×	1.5×	1.4×	1.4×	1.5×
AVCD (Jung et al., 2025)	1.3×	1.6×	1.2×	2.6×	2.7×	3.9×	1.7×	2.1×	5.7×
ASD (Jung et al., 2026)	1.9×	2.4×	1.7×	3.6×	3.5×	5.3×	3.3×	3.3×	5.9×
MAD (Chung et al., 2026)	2.4×	2.7×	1.4×	3.0×	2.9×	2.3×	2.9×	2.8×	2.4×
RED (Ours)	1.6×	1.4×	1.2×	1.4×	1.4×	1.7×	1.7×	1.4×	1.7×

Table F: **Accuracy–TTFT trade-off of SID (Huo et al., 2025) pruning on CMM.** Results with Qwen2.5-Omni-3B under a common 2-fps configuration. Accuracy (%) is reported on all 2,400 questions, and TTFT is normalized to the paired Base. **Bold** indicates the best accuracy or lowest non-base TTFT.

Method	VL	AL	VAL	Vis.	Aud.	Lang.	Avg.	TTFT (s)	Relative
Base model	**90.50**	**84.75**	89.50	69.25	84.50	75.25	82.29	2.717	1.000×
SID (0% removed)	**90.50**	**84.75**	89.50	69.25	84.50	75.25	82.29	4.379	1.612×
SID (60% removed)	87.50	83.75	88.25	71.25	81.25	77.75	81.63	4.179	1.538×
SID (90% removed)	89.25	84.25	91.00	71.25	74.50	**79.00**	81.54	4.101	1.509×
RED (Ours)	88.25	**84.75**	**95.00**	**76.25**	**88.50**	77.25	**85.00**	**3.893**	**1.433×**

B Experimental Details

B.1 Evaluation Protocol

Evaluation. The results in Table 1 are our reproductions using frozen backbones and deterministic decoding. All methods are evaluated on the complete CMM, AVHBench, and SVHalluc benchmarks, with the unweighted mean reported over the six CMM subsets, three AVHBench tasks, and six SVHalluc tasks, respectively. For MUSIC-AVQA, we use all 9,129 questions in the official corrected test annotations; MUSIC-AVQA-Hard contains 2,508 questions. Method-specific decoding settings follow the original papers or released implementations unless stated otherwise.

Parameters and sampling. AVCD (Jung et al., 2025) and MAD (Chung et al., 2026) use 100 randomly sampled labeled examples from each dataset to determine their decoding strengths. **RED** follows the same 100-example setting and randomly samples 100 questions from each benchmark to estimate its score statistics without using answer labels. The resulting statistics are used throughout evaluation on the corresponding complete benchmark. For the remaining baselines, we follow the decoding settings reported in the corresponding papers or released implementations.

B.2 Baselines

We compare RED with two groups of training-free decoding methods: (1) audio-visual hallucination methods, including AVCD (Jung et al., 2025), ASD (Jung et al., 2026), and MAD (Chung et al., 2026); and (2) vision-language hallucination methods, including VCD (Leng et al., 2024), ICD (Wang et al., 2024), CODE (Kim et al., 2024), and SID (Huo et al., 2025). For the second group, we follow the trimodal extension protocol of AVCD (Jung et al., 2025), adapting each method to audio, video, and joint audio-visual inputs while preserving its original decoding rule. For ASD (Jung et al., 2026), originally evaluated for caption generation, we adapt its attention intervention to binary QA as detailed below.

Base model. The pretrained backbones (*i.e.*, VideoLLaMA2.1-7B-AV, Qwen2.5-Omni-3B, and Qwen2.5-Omni-7B) are evaluated using their standard decoding procedures without fine-tuning or additional correction.

VCD (Leng et al., 2024) contrasts the original input with a corrupted counterpart. Following the audio-visual extension introduced in AVCD (Jung et al., 2025), we construct audio, video, and joint corruption branches while retaining the original contrastive decoding rule.

ICD (Wang et al., 2024) obtains a contrastive prediction from a disturbance instruction. We extend it to the audio-visual setting using audio-focused, video-focused, and joint disturbance branches while keeping the corresponding media inputs unchanged.

CODE (Kim et al., 2024) contrasts the original prediction with one conditioned on a self-generated description. For the audio-visual setting, we generate contrastive descriptions covering both audio and video and otherwise retain the released decoding rule.

SID (Huo et al., 2025) constructs a contrastive branch from low-importance perceptual tokens. Following the audio-visual extension used in AVCD (Jung et al., 2025), we select low-importance tokens separately from the audio and video streams while retaining the original decoding rule.

AVCD (Jung et al., 2025) identifies modality dominance and constructs attention-masked contrastive branches accordingly. We follow its original audio-visual formulation and adapt only the modality-token indexing and attention hooks required by each backbone architecture.

ASD (Jung et al., 2026). ASD shifts attention from unimodal toward cross-modal sink tokens and combines the original and recalibrated predictions with an adaptive weight. We retain the original intervention and apply it during answer generation for binary QA. For VideoLLaMA2.1-7B-AV, sink tokens are identified separately within the audio and video streams, with the corresponding dimensions and threshold re-estimated from its activations. We use the official implementation and authors' hyperparameters without additional QA-specific tuning. On Qwen2.5-Omni-7B, the adaptive gate affects 37% of examples and changes 15% of predictions; setting $\alpha = 0$ exactly recovers the base outputs.

MAD (Chung et al., 2026) estimates which modalities are required by the question and uses the resulting relevance scores to combine modality-conditioned predictions. We follow its original audio-visual modality assessment and adaptive decoding rule for each backbone.

B.3 Benchmark details and metrics

CMM. The Curse of Multi-Modalities (CMM) (Leng et al., 2025) evaluates hallucinations across language, visual, and audio modalities through binary object- and event-level probing. It contains six subsets organized by two hallucination sources. Visual-Language (VL), Audio-Language (AL), and Visual-Audio-Language (VAL) evaluate spurious inter-modality correlations, while Visual Dominance (Vis.), Audio Dominance (Aud.), and Language Dominance (Lang.) evaluate overreliance on unimodal priors. Each subset contains 200 samples, and each sample is paired with one positive and one negative probing question, resulting in 1,200 samples and 2,400 Yes/No questions in total. We evaluate the complete benchmark and report the unweighted mean accuracy over the six subsets. For compactness, Table 1 reports the three modality-dominance subsets as Vis., Aud., and Lang., together with the six-subset average; full per-subset results are provided in Table H.

AVHBench. AVHBench (Kim et al., 2025) evaluates cross-modal hallucinations in audio-visual understanding. The full benchmark contains three binary judgment tasks and an audio-visual captioning task. We evaluate the three judgment tasks: Audio-driven Video Hallucination (ADV), Video-driven Audio Hallucination (VDA), and Audio-Visual Matching (AVM). ADV tests whether audio induces hallucinated visual content, VDA tests whether visual content induces hallucinated audio, and AVM evaluates whether the audio and visual streams correspond. The released evaluation set contains 5,302 binary judgment questions over 2,136 videos. We report the unweighted mean accuracy over the three judgment tasks.

SVHalluc. SVHalluc (Zhang et al., 2026) evaluates speech–vision hallucinations along semantic and temporal dimensions. The semantic tasks are Global Semantic Alignment (GSA), Fine-

Grained Semantic Alignment (FGSA), and Cross-Modal Semantic Binding (CMSB). The temporal tasks are Temporal Alignment (TA), Temporal Forecasting (TF), and Cross-Modal Temporal Binding (CMTB). GSA, FGSA, CMSB, and TA use two-choice questions, while TF and CMTB use multiple-choice questions. SVHalluc is constructed from the validation split of YouCook2 and contains 2,405 video-question pairs: 444 GSA, 444 FGSA, 534 CMSB, 368 TA, 276 TF, and 339 CMTB questions. All models are evaluated in a zero-shot setting without additional training. In Table 1, Sem. and Temp. denote the unweighted means over the three semantic and three temporal tasks, respectively, and Avg. denotes the unweighted mean over all six tasks. Full task-level results are provided in Table I.

MUSIC-AVQA. MUSIC-AVQA (Li et al., 2022) evaluates general audio-visual question answering in musical performance videos. The original release contains 9,288 videos spanning 22 instrument categories and more than 150 hours of video, with 45,867 question-answer pairs constructed from 33 question templates. The benchmark covers nine question types across audio, visual, and joint audio-visual settings, including counting, comparison, location, existence, and temporal reasoning. MUSIC-AVQA is formulated as open-ended question answering with 42 possible answer categories. The original dataset contains 32,087 training, 4,595 validation, and 9,185 test question-answer pairs. We use the official corrected test annotations, which contain 9,129 questions, and report overall accuracy on this complete corrected split.[3] The evidence-selection distribution in Figure E uses the same 9,129 questions.

MUSIC-AVQA-Hard. MUSIC-AVQA-Hard (Kim & Seo, 2026) is an audio-sensitive subset of the MUSIC-AVQA test split. It is constructed using a single-frame shortcut test in which GPT-4o receives only the temporally central video frame, without audio or additional frames, in two independent runs. Questions answered correctly in both runs are removed, leaving examples that are less susceptible to vision-only shortcuts. The subset preserves the original MUSIC-AVQA question and answer format. Our evaluation uses 2,508 questions from this filtered subset and reports overall accuracy. All 2,508 questions are included in the evidence-selection distribution.

Metrics. We report accuracy as the primary metric across all benchmarks. For CMM, Avg. is the unweighted mean over all six subsets; Table 1 also reports the three modality-dominance subsets individually. For AVHBench, Avg. is the unweighted mean over its three judgment tasks. For SVHalluc, Sem. and Temp. are the unweighted means over the three semantic and three temporal tasks, respectively, and Avg. is the unweighted mean over all six tasks. For MUSIC-AVQA and MUSIC-AVQA-Hard, we report accuracy over all 9,129 and 2,508 questions in the corresponding evaluation splits, respectively. For analyses based on continuous binary answer scores, we additionally report AUROC to measure discrimination independently of a fixed decision threshold.

B.4 Evidence-intervention protocol and results

A correct prediction does not necessarily imply that the model relied on the evidence required by the question (Xiao et al., 2024; Yin et al., 2025). Following intervention-based diagnostics (Feder et al., 2022; Ribeiro et al., 2020; Kaushik et al., 2020; Gardner et al., 2020), we evaluate 3,359 retained intervention pairs from AVHBench (Kim et al., 2025): 1,621 answer-changing pairs and 1,738 answer-preserving pairs. Sensitivity measures the fraction of answer-changing pairs for which both the original and intervened inputs are answered correctly, whereas invariance violation measures prediction changes under answer-preserving non-required-modality swaps. Within each test, all methods use the same retained pairs and matched donor assignments.

Annotation-guided pair construction. We construct original–contrastive pairs from the Video-driven Audio Hallucination and Audio-driven Video Hallucination tasks in AVHBench (Kim et al., 2025), whose questions require audio and visual evidence, respectively. Each pair retains the question and replaces one modality with the corresponding stream from a donor clip. For sensitivity, we select donors with explicit opposite-label annotations for the same queried entity and required modality. A Yes-to-No replacement is supported by the donor's No annotation, and a No-to-Yes

[3] Official corrected annotations: https://github.com/GeWu-Lab/MUSIC-AVQA/blob/main/data/json_update/avqa-test.json.

Table G: **Construction of evidence interventions.** Required-modality swaps change the answer, whereas irrelevant-modality swaps preserve it. Pairs denote retained interventions and Samples denote unique source video–question items; the total is deduplicated.

Test	Req.	Swap	Gold	Donor	Pairs	Samples
Sens.	Audio	A	Yes→No	Annotated	804	804
			No→Yes	Annotated	524	524
	Video	V	Yes→No	Annotated	192	192
			No→Yes	Annotated	101	101
Inv.	Video	A	unchanged	Random	674	674
	Audio	V	unchanged	Random	1,064	1,064
Total (unique source samples)					3,359	2,680

replacement by its Yes annotation. We retain the question and the original stream of the other modality and assign the donor's annotated answer to the replacement input. For invariance, we retain the question and the required source modality while replacing the non-required modality: video for audio-required questions and audio for video-required questions. The answer remains unchanged because its determining evidence is retained. The evaluation comprises 1,621 sensitivity pairs and 1,738 invariance pairs from 2,680 distinct source video–question items (Table G). The sensitivity set contains 804/524 Yes-to-No/No-to-Yes pairs for audio-required questions and 192/101 for video-required questions; the invariance set contains 1,064 and 674 pairs, respectively.

Annotation and media verification. We cross-reference the original and donor QA records, replacement-media identifiers, and saved predictions to verify the queried entity, answer transition, and stream provenance. Video packet payloads are compared with the expected original or donor source, and the required-modality source paths and donor assignments are checked across evaluation records. For video replacements, we check the source and sample count for the retained, re-encoded audio. We further examine the original and replacement evidence in sensitivity pairs using PANNs (Kong et al., 2020) for audio and OWLv2 (Minderer et al., 2023) for sampled video frames. We use the resulting event and object scores to check consistency with the queried entity and the annotation-defined answer transition. This provides complementary checks and identifies evidence that requires review. AVHBench annotations determine the reference answers. All methods are compared on the same retained pairs with matched donor assignments. For invariance, we use pairs with valid binary responses across methods and backbones.

Pair-level evaluation. Sensitivity requires both the original and replacement answers to be correct. Since their reference answers differ, consistently predicting either Yes or No cannot pass a pair: the model must answer the original correctly and update its answer according to the replacement evidence. We average the Yes-to-No and No-to-Yes success rates equally so that both transition directions contribute equally despite their different sample counts; an invalid response counts as a failure. Invariance violation is the fraction of answer-preserving pairs whose predictions change. For the Target-only reference, only the question-required modality is provided. Replacing a non-required modality therefore leaves the input unchanged, so the same prediction applies to both members, yielding zero invariance violation by construction. Figure F illustrates the pair-level sensitivity criterion: after replacing the required evidence while retaining the other modality, both the original and intervened inputs must be answered correctly. Figure G reports the modality-specific results, while Figure 3 reports their corresponding averages over the two required modalities.

Evidence following under intervention. Figure G examines whether the accuracy gains are accompanied by stronger dependence on question-required evidence. On audio-required questions, **RED** increases sensitivity from 28.0% for Base to 58.3%, while reducing invariance violation from 26.3% to 9.2%. On video-required questions, sensitivity increases from 60.0% to 70.1%, while invariance violation decreases from 7.7% to 3.9%. Among methods using both modalities, **RED** achieves the highest sensitivity for both evidence types and the lowest invariance violation on audio-required questions. For video-required questions, MAD (Chung et al., 2026) achieves a slightly lower invariance violation than **RED** (3.5% vs. 3.9%).

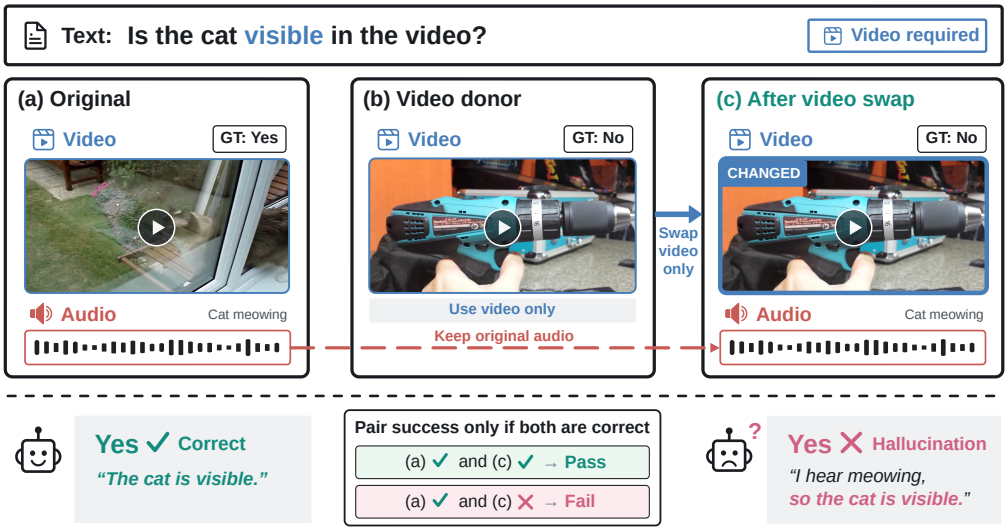


Figure F: **Illustration of pair-level evidence intervention.** For a video-required question, we replace the video of AVHBench sample 01161 with that of sample 00640 while retaining the original audio from 01161. The reference answer changes accordingly, and the pair succeeds only when both the original and intervened inputs are answered correctly. This is an illustrative pair, not a retained evaluation pair; waveforms are schematic.

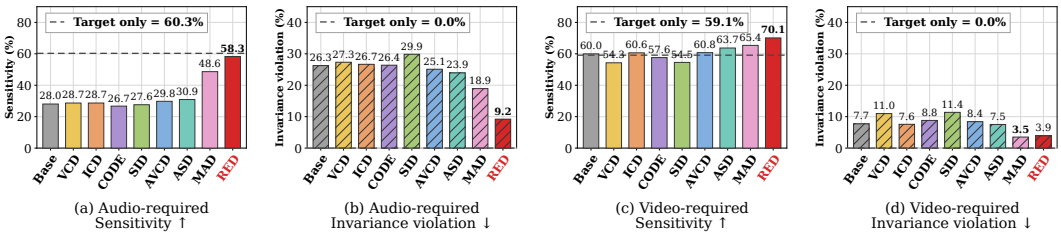


Figure G: **Evidence following under audio-visual interventions on AVHBench.** Sensitivity requires both answers in an answer-changing pair to be correct, while invariance violation measures prediction changes under non-required-modality swaps. Bars and whiskers denote the mean and min–max range across three backbones.

C Additional Experiments

C.1 Detailed CMM and SVHalluc Results

Tables H and I provide subset- and task-level results corresponding to the aggregate CMM and SVHalluc results reported in the main text. **RED** achieves the highest average accuracy across all three backbones on both benchmarks.

The detailed CMM breakdown shows that the gains are concentrated on the modality-dominance subsets: **RED** is best or tied best on Vis. and Aud. across all three backbones. On Qwen2.5-Omni-7B, **RED** achieves absolute gains of 16.5% and 17.3% over the Base model on Vis. and Aud., respectively, while the changes on VL, AL, and VAL are smaller. This pattern further suggests that **RED** is particularly effective when one modality can dominate the prediction.

The task-level breakdown shows that gains are uneven across SVHalluc: CMSB improves over the Base model for all three backbones, whereas temporal changes are more backbone-dependent. For VideoLLaMA2.1-7B-AV, the near-chance GSA and TA results are consistent with the failure pattern reported in SVHalluc (Zhang et al., 2026); for GSA, the benchmark attributes this behavior to an alignment bias toward positive correspondence. Existing decoding baselines largely preserve this pattern.

Table H: **Detailed performance on CMM.** Accuracy (%) on all six CMM subsets; higher is better. VL, AL, and VAL denote Visual-Language, Audio-Language, and Visual-Audio-Language, respectively. Vis., Aud., and Lang. denote Visual Dominance, Audio Dominance, and Language Dominance, respectively. Avg. is the unweighted mean over all six subsets. **Bold** indicates the best result within each backbone.

Method	VL	AL	VAL	Vis.	Aud.	Lang.	Avg.
VideoLLaMA2.1-7B-AV							
Base model	**93.5**	81.2	92.2	73.5	84.0	72.5	82.8
VCD (Leng et al., 2024)	92.5	77.5	89.8	70.8	84.5	73.8	81.5
ICD (Wang et al., 2024)	93.0	81.0	91.5	73.5	84.2	72.5	82.6
CODE (Kim et al., 2024)	92.8	81.8	91.2	72.5	84.0	72.0	82.4
SID (Huo et al., 2025)	90.2	75.0	86.2	65.8	83.5	73.0	79.0
AVCD (Jung et al., 2025)	93.2	77.8	91.5	73.5	82.5	70.8	81.5
ASD (Jung et al., 2026)	92.5	78.8	91.2	74.5	85.5	72.0	82.4
MAD (Chung et al., 2026)	92.5	80.8	**92.5**	82.0	82.8	**77.8**	84.7
RED (Ours)	91.8	**83.2**	91.0	**83.5**	**88.0**	76.2	**85.6**
Qwen2.5-Omni-3B							
Base model	90.5	**87.0**	92.2	69.2	84.2	76.5	83.3
VCD (Leng et al., 2024)	90.0	85.5	93.0	73.2	82.0	76.2	83.3
ICD (Wang et al., 2024)	90.5	85.0	91.8	68.5	84.0	75.8	82.6
CODE (Kim et al., 2024)	91.5	84.5	92.8	69.5	85.2	77.2	83.5
SID (Huo et al., 2025)	91.2	85.0	91.8	71.2	78.0	77.0	82.4
AVCD (Jung et al., 2025)	89.5	85.0	92.2	70.0	85.0	75.5	82.9
ASD (Jung et al., 2026)	91.5	86.5	92.8	70.0	84.8	78.0	83.9
MAD (Chung et al., 2026)	91.2	85.5	90.8	74.5	**87.0**	78.8	84.6
RED (Ours)	**92.5**	84.2	**96.0**	**75.8**	**87.0**	**81.5**	**86.2**
Qwen2.5-Omni-7B							
Base model	**94.8**	83.0	92.5	69.5	71.5	82.2	82.2
VCD (Leng et al., 2024)	92.8	84.2	93.2	69.0	68.5	82.5	81.7
ICD (Wang et al., 2024)	93.5	84.2	92.5	68.8	71.2	81.8	82.0
CODE (Kim et al., 2024)	93.8	83.5	91.8	68.0	68.5	82.8	81.4
SID (Huo et al., 2025)	92.0	86.5	91.0	73.5	65.8	**83.5**	82.0
AVCD (Jung et al., 2025)	87.5	84.2	93.0	69.5	72.8	75.2	80.4
ASD (Jung et al., 2026)	94.0	83.8	92.2	70.0	74.2	82.0	82.7
MAD (Chung et al., 2026)	93.8	85.0	94.5	77.0	83.2	83.0	86.1
RED (Ours)	93.5	**88.5**	**95.2**	**86.0**	**88.8**	83.0	**89.2**

C.2 Evidence attenuation and recovery

Figure H summarizes how support for the correct answer changes across question-only, required-modality-only, full audio-visual, and **RED** predictions on AVHBench. For both audio-required and video-required questions, providing the required modality substantially increases the correct-answer probability relative to question-only inference. Under the full audio-visual input, part of this evidence gain is attenuated, while **RED** recovers and further strengthens support for the correct answer while retaining both modalities. Audio only and Video only denote the task-defined required-modality inputs.

C.3 Comparison with PMI-based decoding

Table J compares **RED** with representative PMI-based decoding methods on CMM using Qwen2.5-Omni-3B. **RED** achieves the highest average accuracy among the compared methods.

Table I: **Detailed performance on SVHalluc.** Accuracy (%) on all six SVHalluc tasks; higher is better. GSA, FGSA, and CMSB denote Global Semantic Alignment, Fine-Grained Semantic Alignment, and Cross-Modal Semantic Binding, respectively. TA, TF, and CMTB denote Temporal Alignment, Temporal Forecasting, and Cross-Modal Temporal Binding. Avg. is the unweighted mean over all six tasks. **Bold** indicates the best result within each backbone, including ties.

Method	Sem.			Temp.			Avg.
	GSA	FGSA	CMSB	TA	TF	CMTB	
VideoLLaMA2.1-7B-AV							
Base model	50.0	67.8	71.3	50.0	30.4	**50.1**	53.3
VCD (Leng et al., 2024)	50.0	70.7	73.0	50.0	30.8	**50.1**	54.1
ICD (Wang et al., 2024)	50.0	68.2	71.3	50.0	30.1	**50.1**	53.3
CODE (Kim et al., 2024)	50.0	68.0	71.7	50.0	29.3	**50.1**	53.2
SID (Huo et al., 2025)	50.0	66.2	71.9	50.0	29.3	49.6	52.8
AVCD (Jung et al., 2025)	50.0	68.5	72.1	50.0	30.4	49.9	53.5
ASD (Jung et al., 2026)	50.0	59.2	70.6	50.0	27.5	48.1	50.9
MAD (Chung et al., 2026)	50.0	75.7	**78.8**	50.0	31.9	49.3	55.9
RED (Ours)	**51.4**	**81.5**	75.1	**52.7**	**33.0**	48.7	**57.1**
Qwen2.5-Omni-3B							
Base model	**74.5**	71.8	75.1	48.1	21.4	49.0	56.7
VCD (Leng et al., 2024)	74.1	70.0	74.2	48.6	24.6	51.3	57.2
ICD (Wang et al., 2024)	74.1	72.3	75.3	48.9	22.1	49.3	57.0
CODE (Kim et al., 2024)	71.6	70.5	74.2	49.2	23.6	49.0	56.3
SID (Huo et al., 2025)	71.2	62.8	70.2	48.1	29.0	52.8	55.7
AVCD (Jung et al., 2025)	73.6	71.8	75.5	50.0	26.4	47.8	57.5
ASD (Jung et al., 2026)	74.3	72.3	74.9	49.2	23.2	49.0	57.1
MAD (Chung et al., 2026)	59.2	**81.8**	84.3	**51.4**	**30.1**	50.1	59.5
RED (Ours)	65.3	80.0	**85.6**	50.0	29.0	**53.4**	**60.5**
Qwen2.5-Omni-7B							
Base model	69.6	83.3	75.5	52.2	33.7	48.7	60.5
VCD (Leng et al., 2024)	70.3	80.2	73.4	**53.3**	34.1	48.1	59.9
ICD (Wang et al., 2024)	69.8	82.4	75.5	51.9	33.7	49.0	60.4
CODE (Kim et al., 2024)	73.0	82.0	74.5	52.4	**35.5**	47.5	60.8
SID (Huo et al., 2025)	70.3	78.4	70.4	53.0	**35.5**	50.1	59.6
AVCD (Jung et al., 2025)	73.0	83.8	76.6	51.4	33.7	**51.9**	61.7
ASD (Jung et al., 2026)	68.0	83.3	74.5	50.3	34.4	50.1	60.1
MAD (Chung et al., 2026)	68.2	**89.6**	86.3	51.6	33.7	49.9	63.2
RED (Ours)	**77.0**	81.3	**87.3**	52.4	33.3	49.0	**63.4**

Table J: **Comparison with PMI-based decoding on CMM.** Accuracy (%) with Qwen2.5-Omni-3B; higher is better.

Method	VL	AL	VAL	Vis.	Aud.	Lang.	Avg.
Base model	90.5	**87.0**	92.2	69.2	84.2	76.5	83.3
CAD (Shi et al., 2024) ($\alpha = 0.5$)	91.2	85.5	93.8	69.2	86.0	80.5	84.4
M3ID (Favero et al., 2024)	91.2	86.2	93.8	69.5	86.0	80.5	84.5
DCPMI (Chen et al., 2026) ($\beta = 0.25$)	**93.0**	85.5	94.2	70.2	86.5	77.2	84.5
RED (Ours)	92.5	84.2	**96.0**	**75.8**	**87.0**	**81.5**	**86.2**

C.4 Detailed MUSIC-AVQA results

Tables K and L provide question-type-level results corresponding to the aggregate MUSIC-AVQA results reported in the main text. **RED** reaches 67.1% overall accuracy on the 9,129-question official corrected MUSIC-AVQA test split and 40.5% on the 2,508-question MUSIC-AVQA-Hard subset.

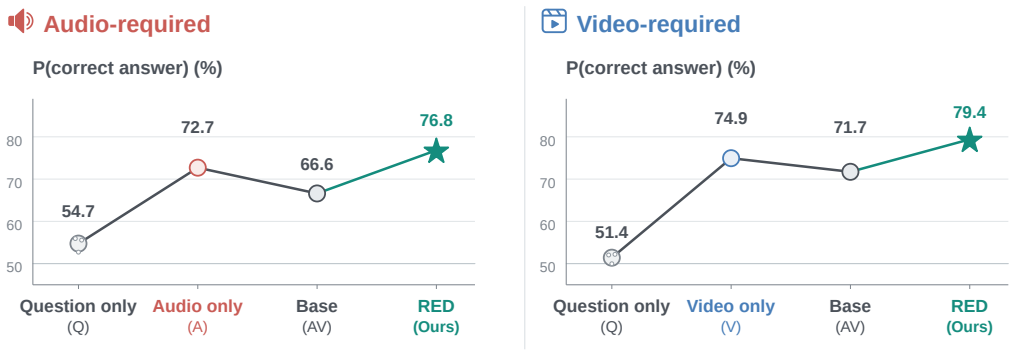


Figure H: **Evidence attenuation and recovery across input conditions.** Correct-answer probability on AVHBench for audio-required and video-required questions, averaged equally across Qwen2.5-Omni-3B, Qwen2.5-Omni-7B, and VideoLLaMA2.1-7B-AV. The required modality provides substantially stronger support than question-only inference; **RED** recovers and further strengthens this support under the full audio-visual input.

Table K: **Performance on MUSIC-AVQA.** Accuracy (%) with Qwen2.5-Omni-7B; higher is better. Avg. denotes accuracy over all 9,129 questions in the official corrected test split. **Bold** indicates the best displayed result in each column, including ties. Cnt., Comp., Loc., Exist., and Temp. denote counting, comparative, location, existential, and temporal questions, respectively.

Method	Audio		Visual		Audio-Visual					
	Cnt.	Comp.	Cnt.	Loc.	Exist.	Loc.	Cnt.	Comp.	Temp.	**Avg.**
Base model	78.6	59.3	71.9	84.2	75.5	47.4	54.2	57.0	43.1	64.5
VCD (Leng et al., 2024)	82.4	64.8	72.2	82.0	74.9	42.8	54.0	58.8	40.5	64.5
ICD (Wang et al., 2024)	82.3	63.6	72.2	84.1	75.6	44.6	56.6	**58.9**	44.3	65.7
CODE (Kim et al., 2024)	80.3	63.5	70.4	83.3	74.5	44.2	48.9	**58.9**	42.8	63.7
SID (Huo et al., 2025)	**84.0**	61.1	72.9	80.2	72.9	42.9	51.1	54.9	36.9	62.9
AVCD (Jung et al., 2025)	81.7	63.8	72.2	82.9	75.4	42.2	54.5	58.7	41.7	64.6
ASD (Jung et al., 2026)	82.1	63.3	72.3	79.5	70.3	38.3	51.2	56.9	39.8	62.4
MAD (Chung et al., 2026)	83.6	64.5	**73.0**	**86.8**	66.4	50.4	**62.6**	57.9	**49.4**	**67.1**
RED (Ours)	79.7	**65.2**	68.1	82.1	**80.1**	**60.5**	57.2	**58.9**	47.3	**67.1**

D Qualitative Analysis

D.1 Evidence Attenuation and Recovery

Figure I provides individual examples of the evidence-attenuation pattern summarized in Figure H. Across one audio-required and two video-required AVHBench cases, the required modality alone provides strong support for the correct answer, while **RED** restores and strengthens this support under the full audio-visual input. These examples illustrate how question-required evidence is recovered at the individual-example level.

D.2 Representative Hallucination Cases

Figure J presents representative audio, visual, and alignment hallucination cases with Qwen2.5-Omni-3B. Across these examples, **RED** corrects predictions that are inconsistent with the question-required audio, video, or audio-visual relation while retaining both perceptual streams.

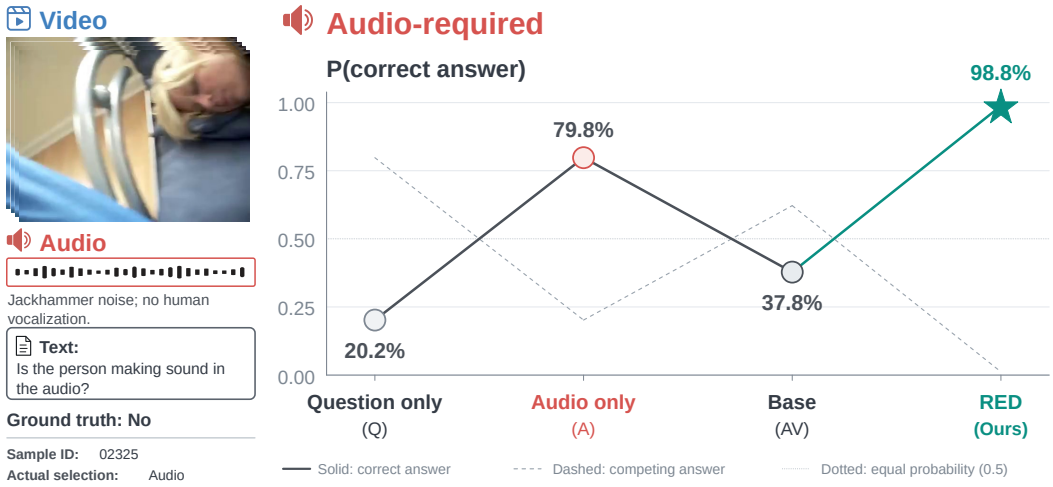
E Limitations

Our current evaluation focuses on audio-visual question answering, including binary, multiple-choice, and open-ended QA, and does not cover long-form generation tasks such as audio-visual captioning (Shen et al., 2023; Zhao et al., 2026a). In addition, **RED** represents question-required evi-

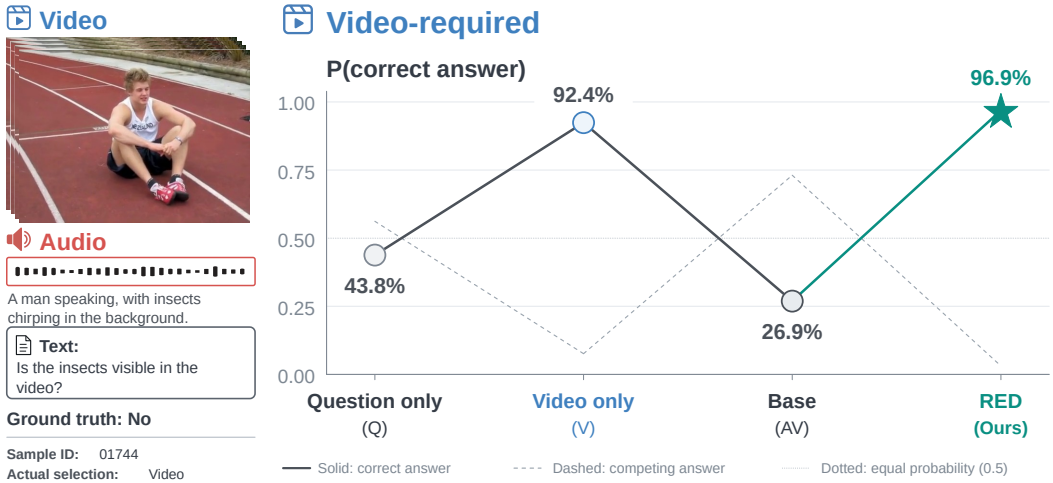
Table L: **Performance on MUSIC-AVQA-Hard.** Accuracy (%) with Qwen2.5-Omni-7B; higher is better. Avg. denotes accuracy over all 2,508 questions. **Bold** indicates the best displayed result in each column, including ties. Cnt., Comp., Loc., Exist., and Temp. denote counting, comparative, location, existential, and temporal questions, respectively.

Method	Audio		Audio-Visual					
	Cnt.	Comp.	Exist.	Loc.	Cnt.	Comp.	Temp.	**Avg.**
Base model	59.7	39.9	43.3	19.6	32.9	31.5	17.5	31.9
VCD (Leng et al., 2024)	**62.3**	41.5	45.5	18.3	33.7	31.5	17.3	32.4
ICD (Wang et al., 2024)	60.2	39.9	43.8	19.1	35.6	31.8	19.1	32.7
CODE (Kim et al., 2024)	59.7	39.4	42.7	19.9	30.0	34.7	19.1	32.3
SID (Huo et al., 2025)	59.7	39.4	49.7	19.1	27.5	32.8	14.7	31.5
AVCD (Jung et al., 2025)	60.2	40.4	43.5	18.3	33.4	31.5	16.7	31.7
ASD (Jung et al., 2026)	61.3	39.4	32.3	18.8	30.0	26.6	14.5	28.2
MAD (Chung et al., 2026)	60.7	44.6	41.0	22.5	41.1	**37.7**	22.8	36.0
RED (Ours)	47.1	**50.3**	**55.3**	**41.6**	**42.1**	36.7	**23.9**	**40.5**

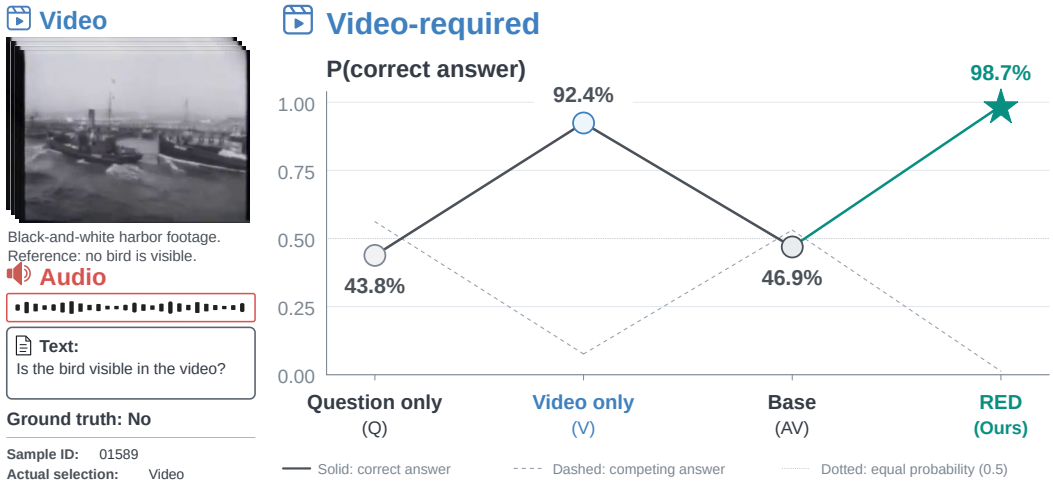
dence using five discrete audio-visual evidence types; while this provides a simple operational mapping between evidence requirements and predictive contributions, finer-grained questions or other modality configurations may require a different evidence space and selection mechanism. Finally, extending **RED** beyond audio-visual settings may require redefining the corresponding contribution decomposition for the modalities involved.



(a) Audio-required: human vocalization.

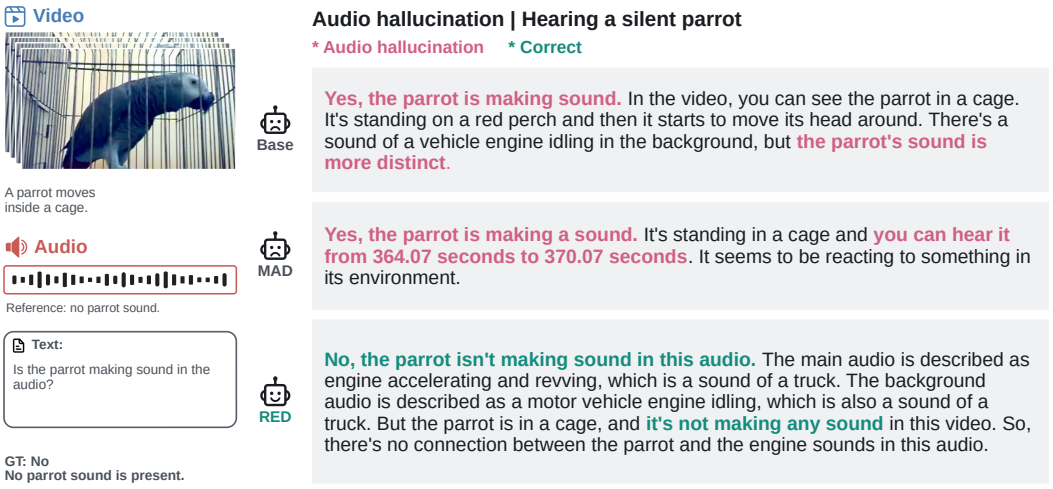


(b) Video-required: insect visibility.

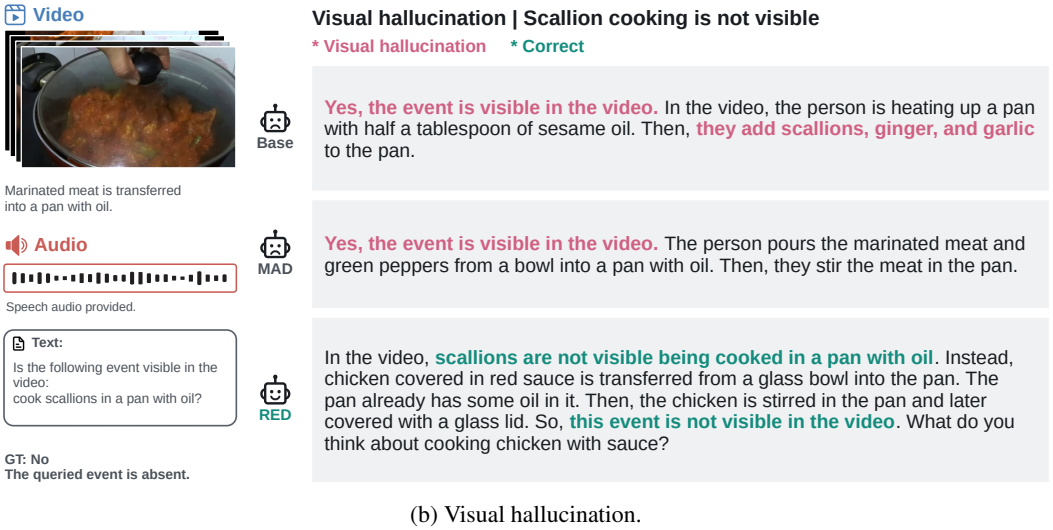


(c) Video-required: bird visibility.

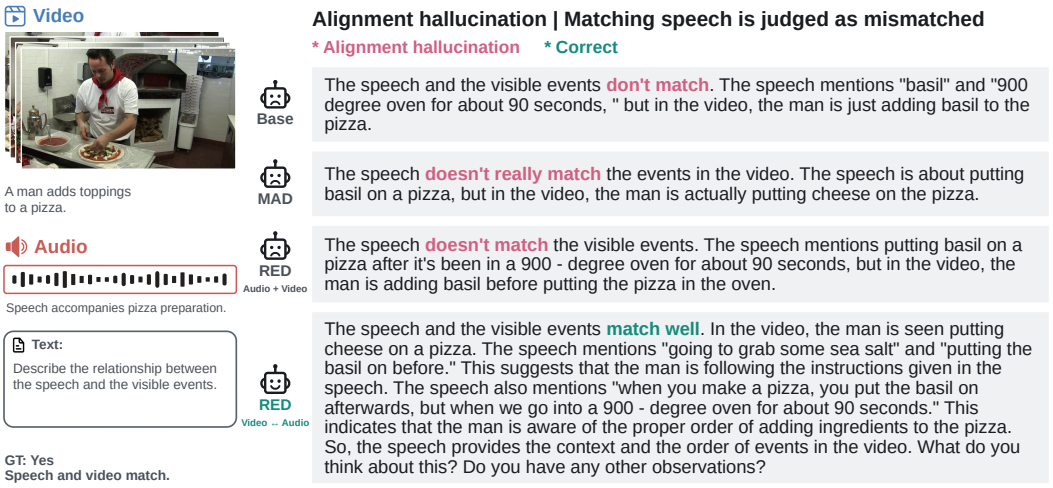
Figure I: **Examples of evidence attenuation and recovery with Qwen2.5-Omni-7B.** Representative AVHBench cases where **RED** recovers question-required evidence under full audio-visual input.



(a) Audio hallucination.



(b) Visual hallucination.



(c) Alignment hallucination.

Figure J: **Qualitative examples with Qwen2.5-Omni-3B.** Representative hallucination cases and corresponding predictions.