CAD: CONFLICT-AWARE DECODING TO MITIGATE CROSS-MODAL HALLUCINATIONS IN OMNIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Omnimodal large language models (Omni-LLMs) integrate audio, video, and text for perception and reasoning but remain vulnerable to cross-modal hallucinations, where one modality improperly influences predictions about another. By avoiding the additional parameter updates required by training-based methods, training-free decoding offers an appealing approach to mitigating these hallucinations. Existing training-free methods modulate modality contributions during decoding through perturbation or relevance weighting, but do not explicitly assess fusion effects on joint predictions, potentially retaining harmful interference or suppressing useful complementarity. To this end, we propose Conflict-Aware Decoding (CAD), a training-free, two-stage framework. First, to assess the potential effects of cross-modal fusion, Potential Conflict Magnitude Estimation (PCME) quantifies audio-video disagreement and the joint prediction's deviation from a relevance-weighted unimodal reference. Second, to assess evidential support for intervention, Conflict Actionability Assessment (CAA) builds on Dempster-Shafer reliability discounting to evaluate task-space answer relations using query relevance and answer decisiveness. When warranted, CAD shifts decoding weight from the joint branch to the unimodal branches. Extensive experiments demonstrate that CAD outperforms competitive training-free methods. It achieves substantial gains on cross-modal hallucination benchmarks CMM and AVHBench while also improving general audio-visual question answering on WorldSense and VideoMME.

1 INTRODUCTION

Omnimodal large language models (Omni-LLMs) extend the capabilities of multimodal LLMs by jointly modeling text, vision, and audio within a unified framework for multimodal perception and reasoning. Omni-LLMs are now widely applied to complex tasks (Geng et al., 2025; Liu et al., 2024) such as audio-visual scene understanding and interactive assistance.

Despite significant progress, Omni-LLMs still face a critical challenge: cross-modal hallucination, which substantially undermines their reliability and stability in practical applications. Effective cross-modal fusion can exploit complementary audio and visual evidence, whereas poorly coordinated fusion may introduce harmful interference. Unlike single-modal hallucinations, cross-modal hallucinations reflect inadequate control over modality interactions, whereby information from one modality improperly influences the generation of content associated with another. Visual and acoustic cues may induce descriptions of nonexistent sounds and visual events, respectively.

Recent decoding methods mitigate multimodal hallucinations by modulating the influence of different modalities during generation. Audio-Visual Contrastive Decoding (AVCD) uses attention-based modality dominance to guide attentive masking and generate perturbed logits for contrastive decoding (Jung et al., 2025). Modality-Adaptive Decoding (MAD) estimates query-conditioned modality relevance and adaptively combines the audio-only, video-only, joint audio-visual, and text-only decoding branches (Chung et al., 2026). However, as shown in Fig. 1(a), neither attentive masking nor relevance-based branch weighting explicitly assesses predictive compatibility between joint and unimodal branches to guide intervention. Consequently, these methods lack an explicit criterion for deciding whether fusion discrepancies warrant suppression or reflect useful complementarity.

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

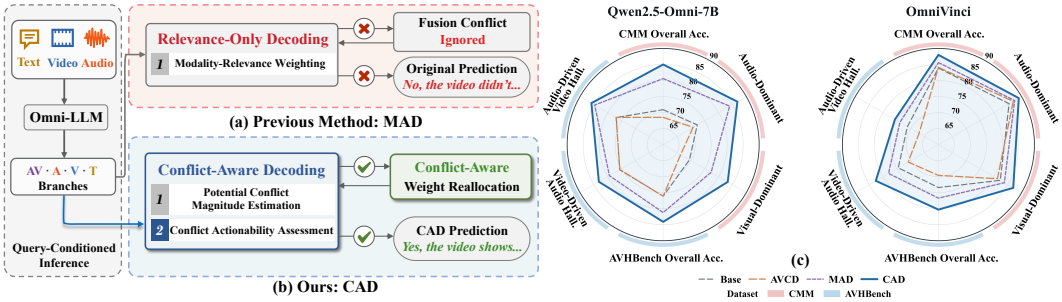

Figure 1: (a) MAD relies on modality relevance alone, overlooking cross-modal conflicts. (b) CAD assesses conflict magnitude and actionability, then reallocates decoding weights. (c) On CMM and AVHBench, CAD consistently outperforms baselines with both Qwen2.5-Omni-7B and OmniVinci.

To examine fusion outcomes, we evaluate Qwen2.5-Omni-7B on 800 CMM (Leng et al., 2025) questions from its Visual Dom. and Audio Dom. categories under audio-only, video-only, and joint audio-visual inputs. Joint inference is incorrect in 4.5% of cases where both unimodal branches are correct, yet correct in 8.7% of cases where both are incorrect. Predictive disagreement may reflect harmful interference or useful complementarity, motivating us to assess both discrepancy magnitude and intervention support beyond modality relevance.

To this end, we propose **Conflict-Aware Decoding (CAD)**, a training-free framework that augments relevance-aware decoding with a two-stage conflict assessment. As illustrated in Fig. 1(b), **Potential Conflict Magnitude Estimation (PCME)** introduces the unsigned score C to measure the magnitude of potential cross-modal conflict. Specifically, PCME uses normalized Jensen-Shannon divergence to quantify disagreement between unimodal predictions and the joint prediction's deviation from their relevance-weighted mixture. Moreover, **Conflict Actionability Assessment (CAA)** computes a conflict actionability score B by evaluating task-space answer relations using query relevance and answer decisiveness as reliability evidence. Following reliability discounting in Dempster-Shafer (Bundy & Wallen, 1984) evidence theory, CAA represents committed support and ignorance through simple-support masses. When C reaches or exceeds the intervention threshold and B provides evidence for intervention, CAD shifts weight from the joint branch to the unimodal branches and combines all four branch logits using the adjusted weights for conflict-aware generation.

Extensive experiments on CMM, AVHBench, WorldSense and VideoMME show that CAD achieves consistent gains over baselines across four Omni-LLMs. As shown in Fig. 1(c), CAD improves overall accuracy over MAD on CMM and AVHBench by 4.4 and 2.9 percentage points with Qwen2.5-Omni-7B, respectively, and by 2.4 and 3.6 points with OmniVinci.

We summarize the contributions of this paper as follows.

- Motivated by the effects of cross-modal fusion on joint predictions, we propose CAD, a training-free, two-stage framework that augments relevance-aware decoding with conflict assessment and reweighting to mitigate cross-modal hallucinations in Omni-LLMs.
- We introduce PCME and CAA to assess potential conflict magnitude and actionability. PCME quantifies predictive discrepancies across modalities to characterize potential conflict, while CAA draws on Dempster-Shafer reliability discounting to assess task-space evidence for selective weight reallocation.
- Experiments demonstrate that CAD consistently outperforms baselines on the cross-modal hallucination benchmarks CMM and AVHBench across multiple backbones, with additional gains on the general audio-visual benchmarks WorldSense and VideoMME.

2 Related Work

2.1 Omnimodal Large Language Models

In recent years, multimodal large language models (MLLMs) have evolved from static image-text understanding toward multimodal reasoning in complex audio-visual scenarios. Conventional

MLLMs (Bai et al., 2025; Chen et al., 2024; Li et al., 2025; Lin et al., 2024; Sellergren et al., 2026; Sun et al., 2024) typically process different modalities relatively independently. In contrast, omnimodal large language models (Omni-LLMs) (Fu et al., 2025b; Li et al., 2024; Xu et al., 2025b; Team, 2026), such as Qwen-Omni and OmniVinci (Xu et al., 2025a; Ye et al., 2025), project text, image, video, and audio features into a unified representation space, while organizing audio and visual tokens along a shared timeline into an interleaved sequence (AI et al., 2025; Tong et al., 2025; Xie & Wu, 2024). Effective reasoning over this interleaved sequence requires temporal modeling and complementary evidence integration, but inconsistent signals may lead to harmful fusion conflicts.

2.2 Mitigating Hallucinations in MLLMs

Hallucination remains a critical challenge for MLLMs, whose responses may deviate from evidence in audio, visual, or textual inputs. Existing mitigation approaches include representation debiasing (Jiang et al., 2024; Zhang et al., 2025), architectural refinement (Wang et al., 2025; Xing et al., 2024; Yang et al., 2025; Huang et al., 2024), and training-free inference. Among them, training-free decoding methods (Li et al., 2023; Cho et al., 2025) are appealing because they require neither additional annotations nor parameter updates. They suppress unsupported generation through modality perturbations (Leng et al., 2024), auxiliary descriptions (Kim et al., 2024), instruction-level contrasts (Wang et al., 2024), or layerwise distribution differences (Chuang et al., 2024). However, these approaches are primarily designed for vision-language settings and therefore do not directly address hallucinations arising from interactions between audio and visual evidence in Omni-LLMs.

To address this limitation, recent studies have extended training-free decoding to audio-visual settings. AVCD (Jung et al., 2025) uses attention-derived dominance scores to identify less dominant modalities and applies attentive masking to construct contrastive logits over audio, visual, and textual inputs. MAD (Chung et al., 2026) instead estimates query-specific modality relevance through model self-assessment and uses the scores to adaptively weight modality-specific decoding branches. Despite these advances, these methods do not explicitly assess whether cross-modal fusion integrates complementary evidence or introduces harmful interference. Differently, CAD models cross-modal fusion, aiming to reduce harmful interference while preserving useful cross-modal complementarity.

3 Preliminaries

3.1 Modality-Adaptive Decoding

Modality-Adaptive Decoding (MAD) is a training-free method that extends contrastive decoding to audio-visual inputs. It uses query-specific modality relevance to determine how strongly to contrast predictions with and without a modality.

Given audio input X_A, video X_V, and textual query X_Q, MAD considers four branches $\mathcal{K} = \{AV, A, V, T\}$, with contexts $X^{(AV)} = (X_A, X_V)$, $X^{(A)} = X_A$, $X^{(V)} = X_V$, and $X^{(T)} = \varnothing$. All branches share model parameters θ, query X_Q, and generated prefix $y_{<t}$; the question-only branch T omits audio and video but retains the query. At step t, the logit for token v in vocabulary $\mathcal{V}$ is

$$z_k^t(v) = \mathrm{logit}_\theta\Big(v \mid X^{(k)}, X_Q, y_{<t}\Big), \qquad k \in \mathcal{K}, \quad v \in \mathcal{V}. \tag{1}$$

To estimate modality relevance, MAD queries the same model with (X_A, X_V, X_Q) and an additional prompt X_M asking whether answering the query requires audio, video, or both. Let u_A, u_V, and u_{AV} denote the next-token logits for the responses "audio," "video," and "both," respectively. Normalizing these three logits gives the category-level relevance scores

$$(r_A, r_V, r_{AV}) = \mathrm{softmax}([u_A, u_V, u_{AV}])\,. \tag{2}$$

These scores are computed once before answer generation and remain fixed throughout the response. MAD uses these scores to scale four contrastive terms. For example, contrasting A against T yields $z_A^t(v) + \gamma r_A[z_A^t(v) - z_T^t(v)]$, where $\gamma > 0$ controls the overall contrastive strength. This term adds the audio-induced logit difference, scaled by γr_A, to the audio-only logits. MAD similarly contrasts AV against A and V, each with strength γr_{AV}, and V against T with strength γr_V. The sum of

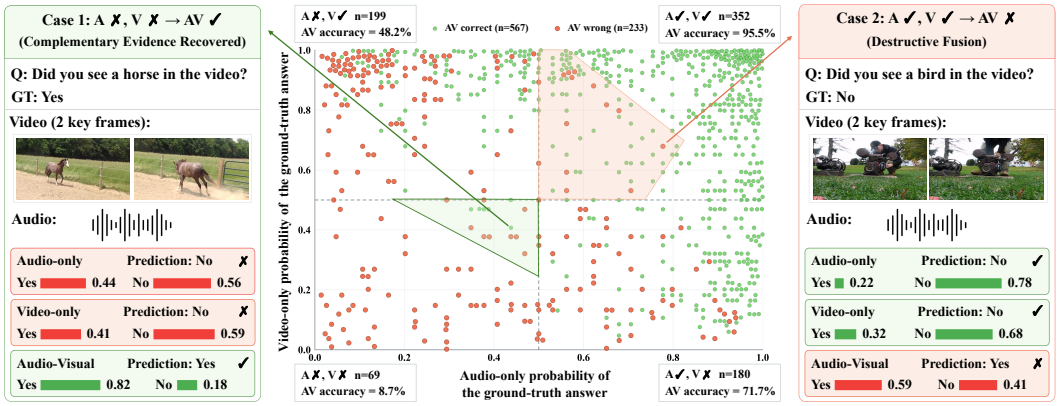

Figure 2: Fusion outcomes of Qwen2.5-Omni-7B on CMM. Axes show unimodal ground-truth probabilities. Examples show destructive fusion and complementary recovery.

these four contrastive terms gives the MAD logits, which can be written compactly as

$$z^t_{\mathrm{MAD}}(v) = \sum_{k\in\mathcal{K}} w_k z^t_k(v),$$
$$w_{AV} = 2 + 2\gamma r_{AV}, \qquad w_A = 1 + \gamma(r_A - r_{AV}),$$
$$w_V = 1 + \gamma(r_V - r_{AV}), \qquad w_T = -\gamma(r_A + r_V). \tag{3}$$

The w_k are signed logit coefficients, not mixture probabilities, and satisfy $\sum_{k\in\mathcal{K}} w_k = 4$. During generation, coefficients remain fixed, while branch logits change with the shared prefix. MAD decodes via $\hat{y}_t = \arg\max_{v\in\mathcal{V}} z^t_{\mathrm{MAD}}(v)$. CAD builds on these predictions and coefficients.

3.2 Motivation

MAD uses query-specific modality relevance to set the branch coefficients in equation 3. Relevance indicates which modalities a query requires, but does not directly establish the reliability of the resulting joint prediction.

To examine how fusion affects predictions, we compare the predictions of the A, V, and AV branches of Qwen2.5-Omni-7B on 800 CMM questions, judging correctness against the ground-truth answers. Figure 2 shows that AV is incorrect in 4.5% of cases where both unimodal branches are correct, yet correct in 8.7% of cases where both are incorrect. The highlighted cases illustrate opposite fusion outcomes despite identical cross-branch answer relations.

These outcomes motivate assessing predictive discrepancies in addition to modality relevance. Quantifying their magnitude characterizes potential conflict at the distribution level, but does not establish whether intervention is warranted: disagreement may reflect either harmful interference or useful complementarity. Therefore, CAD combines magnitude estimation with an assessment of intervention support to guide selective weight reallocation.

4 Method

In this section, we introduce **Conflict-Aware Decoding (CAD)**. CAD extends MAD's relevance-aware decoding with conflict-aware intervention during generation, aiming to reduce harmful interference from audio-visual fusion while preserving useful cross-modal complementarity.

4.1 Overview of CAD

As illustrated in Fig. 3, CAD is a training-free, two-stage conflict-aware decoding method for mitigating audio-visual cross-modal hallucinations in Omni-LLMs during inference.

In the first stage, CAD performs **Potential Conflict Magnitude Estimation (PCME)** and introduces an unsigned score C that measures the magnitude of potential cross-modal conflict. C cap-

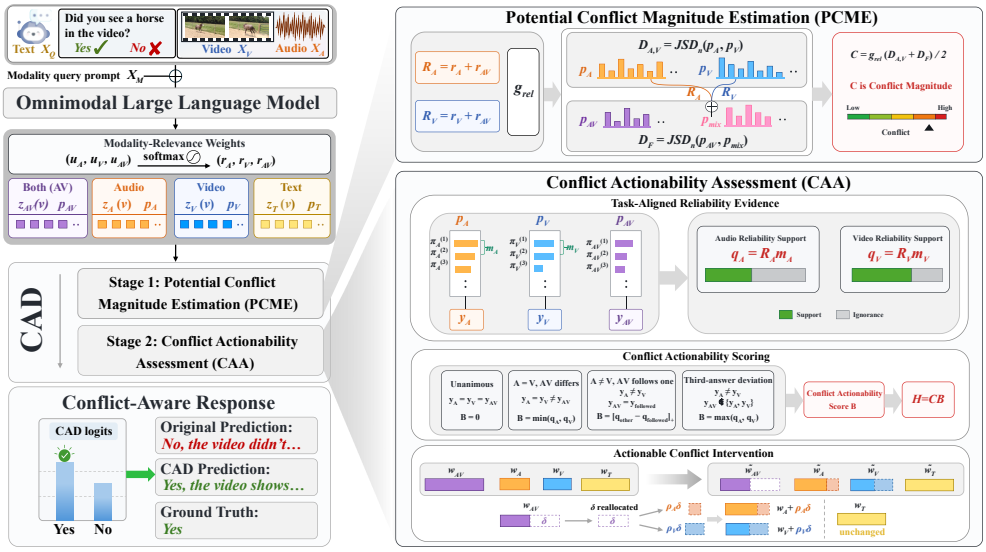

Figure 3: Overview of CAD. PCME estimates potential conflict magnitude C, and CAA assesses intervention support B. Their product H guides selective reallocation of decoding weights.

tures two forms of predictive discrepancy: (i) disagreement between the audio-only and video-only predictions; and (ii) the joint audio-visual prediction's deviation from a reference distribution that combines the two unimodal predictions using query-specific relevance weights.

In the second stage, CAD performs **Conflict Actionability Assessment (CAA)** and introduces an actionability score B that quantifies the evidential support for intervening on the potential cross-modal conflict indicated in the first stage. Specifically, based on Dempster-Shafer evidence theory, CAD derives modality reliability from query relevance and answer decisiveness. It then combines C and B to identify actionable cross-modal conflicts and selectively adjust decoding weights.

4.2 Potential Conflict Magnitude Estimation

Building on MAD's branch predictions and modality-relevance weights, PCME estimates the magnitude of potential cross-modal conflict by comparing the audio-only, video-only, and joint audio-visual predictions. Because the question-only branch T contains no audio-visual evidence, PCME excludes it from conflict estimation. For each remaining branch $k \in \{AV, A, V\}$, we convert the logit vector defined in equation 1 into the next-token distribution

$$p_k^t(v) = \left[\mathrm{softmax}(z_k^t)\right]_v, \qquad v \in \mathcal{V}. \tag{4}$$

In our implementation, we assess conflict and reallocate coefficients once at the initial answer-generation step, keeping them fixed throughout the response. All distributions and scores below refer to this step, with superscript t omitted. Because the joint-use category involves both modalities, we define the effective audio and video relevance scores as $R_A = r_A + r_{AV}$ and $R_V = r_V + r_{AV}$, respectively. Using these scores, PCME introduces a relevance-balance factor to focus conflict estimation on cases where both modalities are relevant to the query:

$$g_{\mathrm{rel}} = \frac{2\min(R_A, R_V)}{R_A + R_V + \epsilon}, \tag{5}$$

where ϵ is a numerical-stability constant set to 10^{-8}. The factor approaches one when the effective relevance scores are balanced and decreases as either modality becomes less relevant. Because $2\min(R_A, R_V) \le R_A + R_V$ and $\epsilon > 0$, we have $g_{\mathrm{rel}} \in [0, 1)$. PCME then constructs a relevance-weighted unimodal-branch reference:

$$p_{\mathrm{mix}} = \frac{R_A p_A + R_V p_V}{\|R_A p_A + R_V p_V\|_1}. \tag{6}$$

To account for the modalities' differing relevance to the query, PCME uses their weighted mixture p_{mix} as a unimodal reference. Using Jensen–Shannon divergence normalized by $\log 2$, denoted by

$\mathrm{JSD}_n \in [0, 1]$, PCME measures two complementary predictive discrepancies: (i) disagreement between the audio-only and video-only predictions, $D_{A,V} = \mathrm{JSD}_n(p_A, p_V)$; and (ii) deviation of the joint prediction from the relevance-weighted unimodal reference, $D_F = \mathrm{JSD}_n(p_{AV}, p_{\mathrm{mix}})$. PCME then averages the two discrepancies without additional weighting hyperparameters and modulates their mean by the relevance-balance factor, yielding the potential-conflict magnitude:

$$C = g_{\mathrm{rel}} \frac{D_{A,V} + D_F}{2}. \tag{7}$$

The unsigned score $C \in [0, 1)$ measures potential cross-modal conflict magnitude. Predictive discrepancies may reflect either harmful interference or useful complementarity, so a large C alone does not establish that the joint prediction is erroneous. CAD uses $C \geq \tau$ as a necessary condition for intervention, where τ is the intervention threshold. Therefore, CAD further assesses whether the observed predictive discrepancy warrants selective intervention.

4.3 Conflict Actionability Assessment

Task-Aligned Reliability Evidence. PCME measures full-vocabulary predictive discrepancies using Jensen–Shannon divergence. A high C may reflect harmful interference from audio-visual fusion, but can also arise in the following two cases. First, probability may shift substantially among tokens outside the valid answer set while all branches retain the same answer. Second, the joint branch may deviate from unimodal distributions by integrating complementary audio and visual cues to support the correct answer. A large C alone does not justify suppressing the joint prediction. CAA therefore projects the A, V, and AV predictions onto a shared task-answer space $\mathcal{Y}$:

$$\pi_k(y) = \frac{\sum_{v \in \mathcal{T}(y)} p_k(v)}{\sum_{y' \in \mathcal{Y}} \sum_{v \in \mathcal{T}(y')} p_k(v)}, \qquad k \in \{A, V, AV\}, \tag{8}$$

where $\mathcal{T}(y) \subseteq \mathcal{V}$ contains tokenizer-valid single-token surface forms of y. For multiple-choice and binary tasks, $\mathcal{Y}$ contains the valid options or answer labels. CAA summarizes each projected distribution by its predicted answer $y_k = \arg\max_{y \in \mathcal{Y}} \pi_k(y)$ and margin $m_k = \pi_k^{(1)} - \pi_k^{(2)}$, where $\pi_k^{(1)} \geq \pi_k^{(2)}$ are the two largest values of π_k. The margin m_k measures answer decisiveness within the task space, while R_k reflects the relevance of modality k to the query. CAA combines these two signals as $q_A = R_A m_A$ and $q_V = R_V m_V$. Each q_k serves as an uncalibrated reliability proxy rather than the probability that y_k is correct. Following the reliability-discounting principle of Dempster-Shafer theory, CAA uses these proxies to construct the simple-support masses:

$$M_k(\{y_k\}) = q_k, \qquad M_k(\mathcal{Y}) = 1 - q_k, \quad k \in \{A, V\}. \tag{9}$$

The committed mass expresses reliability-discounted support for y_k; the remainder expresses ignorance. Unlike independent sensors, the audio and video branches share the backbone, query, and language prior, so directly applying Dempster's rule may double-count shared evidence. Our conservative rule uses $\min(q_A, q_V)$ when unimodal branches agree but the joint branch differs; when they disagree and the joint branch follows one, it retains only the opposing branch's bounded positive reliability advantage. This preserves the Dempster–Shafer distinction between committed support and ignorance while adapting aggregation to dependent branches, providing operational evidence for intervention. C measures distributional conflict via JSD rather than Dempster's conflict coefficient.

Conflict Actionability Scoring. CAA maps the answer relations and reliability proxies to

$$B = \begin{cases} 0, & y_A = y_V = y_{AV}, \\ \min(q_A, q_V), & y_A = y_V \neq y_{AV}, \\ [q_{\bar{k}} - q_k]_+, & y_A \neq y_V,\ y_{AV} = y_k, \\ \max(q_A, q_V), & y_A \neq y_V,\ y_{AV} \notin \{y_A, y_V\}, \end{cases} \tag{10}$$

where $[x]_+ = \max(x, 0)$. $k \in \{A, V\}$ denotes the unimodal branch followed by AV, and $\bar{k}$ the other branch. When all branches agree, no evidence supports intervention, so $B = 0$ even if their full-vocabulary distributions differ. When the unimodal branches agree but AV differs, intervention is supported only by their shared reliability, yielding the conservative minimum. When the unimodal answers differ and AV follows branch k, support is positive only if the competing branch provides

stronger reliable evidence. When AV produces a third answer, CAA uses the strongest unimodal opposition, $\max(q_A, q_V)$, as a conservative heuristic. This case is not a strict Dempster derivation and does not imply that either unimodal answer is correct. Across all cases, B measures how strongly the observed answer relation warrants intervention; it neither labels the interaction as harmful nor identifies the ground-truth answer. CAA then combines conflict magnitude with evidential support for intervention to form the actionable-conflict signal $H = CB$, where $0 \leq B \leq 1$ and $0 \leq C < 1$.

Actionable Conflict Intervention. Using the coefficients defined in equation 3, let $s = [w_A]_+ + [w_V]_+$. CAD allocates the transferred influence in proportion to the positive unimodal coefficients:

$$(\rho_A, \rho_V) = \begin{cases} ([w_A]_+/s, [w_V]_+/s), & s > 0, \\ (1/2, 1/2), & s = 0. \end{cases} \tag{11}$$

For intervention strength $\lambda > 0$, the transfer amount is $\delta = \lambda H[w_{AV}]_+$ when $C \geq \tau$ and $\delta = 0$ otherwise. The adjusted coefficients are

$$\begin{aligned} \widetilde{w}_{AV} &= w_{AV} - \delta, & \widetilde{w}_A &= w_A + \rho_A \delta, \\ \widetilde{w}_V &= w_V + \rho_V \delta, & \widetilde{w}_T &= w_T. \end{aligned} \tag{12}$$

B modulates the transfer amount; allocation follows positive base coefficients. Since $\rho_A + \rho_V = 1$, reallocation preserves the coefficient sum and w_T, but need not weaken support for the joint branch's preferred answer in the combined prediction. If $C < \tau$ or $B = 0$, $\delta = 0$, recovering MAD. For $C \geq \tau$ and $w_{AV} > 0$, $\lambda H > 1$ makes $\widetilde{w}_{AV} < 0$, yielding contrastive subtraction of the joint branch. At step t, CAD combines branch logits with adjusted coefficients to select the next token:

$$\hat{y}_t = \arg\max_{v \in \mathcal{V}} \sum_{k \in \{AV, A, V, T\}} \widetilde{w}_k z_k^t(v). \tag{13}$$

CAD computes C, B, and the adjusted coefficients from the initial next-token distributions and keeps the coefficients fixed throughout generation.

5 Experiments

5.1 Experimental Setting

Benchmarks. We evaluate our method on four representative benchmarks. CMM (Leng et al., 2025) and AVHBench (Sung-Bin et al., 2025) assess cross-modal hallucinations in audio-visual large language models. CMM evaluates modality dominance and cross-modal interference across diverse audio-visual scenarios, while AVHBench focuses on video-driven audio and audio-driven video hallucinations. WorldSense (Hong et al., 2026) and VideoMME (Fu et al., 2025a) assess general audio-visual question-answering capabilities. WorldSense emphasizes integrating audio and visual evidence across real-world scenarios, while VideoMME covers short, medium, and long videos.

Baselines. We compare CAD with original backbone decoding (Base) and two training-free decoding methods: AVCD (Jung et al., 2025) and MAD (Chung et al., 2026). AVCD uses attention-based modality dominance for attentive masking in contrastive decoding; MAD adaptively weights four branches using self-assessed query-specific modality relevance.

Implementation Details. We use four audio-visual large language models as backbones: VideoLLaMA2-AV (Cheng et al., 2024), OmniVinci (Ye et al., 2025), Qwen2.5-Omni-3B, and Qwen2.5-Omni-7B (Xu et al., 2025a). All experiments are conducted on NVIDIA L20 GPUs with 48 GB of memory. Following prior work, we sample videos from CMM, AVHBench, and VideoMME at 2 fps, with a maximum of 768 frames per video. For WorldSense, we sample at 1 fps with a maximum of 384 frames. In all experiments, we set the decoding temperature to 0 for deterministic generation across all models. The intervention threshold is set to $\tau = 0.05$.

5.2 Main Results

Results on cross-modal hallucination benchmarks. Table 1 compares CAD with Base, AVCD, and MAD on CMM and AVHBench across four backbones. CAD achieves the highest overall accuracy across backbones and benchmarks, matching or exceeding the strongest baseline in each hallucination category. Relative to MAD, CAD improves overall accuracy by 1.3–4.4 percentage points

Table 1: Main results on cross-modal hallucination benchmarks. We compare the original decoding strategy (Base), AVCD, MAD, and our proposed CAD. All results are reported in accuracy (%).

Model	CMM			AVHBench		
	Visual Dom.	Audio Dom.	Overall Acc.	Video-Driven Audio Hall.	Audio-Driven Video Hall.	Overall Acc.
VideoLLaMA2-AV	72.3	79.5	75.9	75.6	78.8	76.6
VideoLLaMA2-AV + AVCD	72.5	81.8	77.1	72.1	76.8	73.6
VideoLLaMA2-AV + MAD	81.8	82.3	82.0	80.3	78.5	79.7
VideoLLaMA2-AV + CAD	**84.5**	**82.3**	**83.4**	**81.3**	**80.6**	**81.1**
OmniVinci	82.3	85.5	83.9	74.5	71.2	73.4
OmniVinci + AVCD	81.3	87.0	84.1	70.9	67.0	69.6
OmniVinci + MAD	83.8	87.5	85.6	78.0	73.9	76.7
OmniVinci + CAD	**87.0**	**89.0**	**88.0**	**82.6**	**75.7**	**80.3**
Qwen2.5-Omni-3B	69.3	82.5	75.9	74.3	80.2	76.2
Qwen2.5-Omni-3B + AVCD	75.3	78.0	76.6	73.5	78.7	75.2
Qwen2.5-Omni-3B + MAD	76.0	87.3	81.6	80.2	82.1	80.8
Qwen2.5-Omni-3B + CAD	**77.3**	**88.5**	**82.9**	**82.6**	**82.2**	**82.5**
Qwen2.5-Omni-7B	69.5	72.3	70.9	75.7	76.9	76.1
Qwen2.5-Omni-7B + AVCD	66.3	70.8	68.5	75.6	77.0	76.1
Qwen2.5-Omni-7B + MAD	77.3	84.0	80.6	79.4	84.7	81.2
Qwen2.5-Omni-7B + CAD	**83.3**	**86.8**	**85.0**	**83.2**	**85.9**	**84.1**

Table 2: Results on general audio-visual benchmarks using VideoLLaMA2-AV. All results are reported in accuracy (%). Bold denotes the best result in each column.

Method	WorldSense									VideoMME			
	Tech & Science	Culture & Politics	Daily Life	Film & TV	Performance	Games	Sports	Music	Avg.	Short	Medium	Long	Avg.
Base	29.8	25.2	22.6	25.6	25.8	24.9	25.1	28.8	25.9	58.1	45.9	39.9	48.0
MAD	28.0	22.7	20.4	25.3	**26.2**	23.2	25.1	30.5	25.0	58.6	46.3	40.6	48.5
CAD	**30.4**	**27.8**	**25.5**	**27.7**	25.8	**25.8**	**28.4**	**38.2**	**28.8**	**59.9**	**47.1**	**41.2**	**49.4**

on CMM and 1.4–3.6 points on AVHBench. Specifically, on Qwen2.5-Omni-7B, CAD reaches 85.0% on CMM and 84.1% on AVHBench, exceeding Base by 14.1 and 8.0 percentage points and MAD by 4.4 and 2.9 points, respectively. These gains highlight CAD's value in assessing predictive compatibility beyond modality relevance for reliable omnimodal decoding.

Results on general audio-visual benchmarks. Table 2 compares CAD with Base and MAD on WorldSense and VideoMME (without subtitles) using VideoLLaMA2-AV. CAD achieves the highest average accuracy on both benchmarks, outperforming MAD in seven WorldSense categories and all VideoMME categories. Specifically, CAD reaches 28.8% on WorldSense and 49.4% on VideoMME, exceeding Base by 2.9 and 1.4 percentage points and MAD by 3.8 and 0.9 points, respectively. These results suggest that CAD's benefits extend beyond hallucination-specific benchmarks to general audio-visual question answering.

5.3 Ablation Study

Comparison of Weighting Strategies. We compare CAD with two simplified variants on AVHBench using Qwen2.5-Omni-3B. Uniform sets $r_A = r_V = r_{AV} = 1/3$, whereas Argmax uses a one-hot assignment to the highest-scoring category. Both decode with the resulting branch coefficients, without conflict assessment or weight reallocation. As shown in Table 3, both outperform Base, while CAD reaches 82.5% overall accuracy, exceeding Uniform and Argmax by 4.5 and 2.9 percentage points, respectively, with gains in both hallucination categories. These results support combining relevance-aware weighting with conflict-aware reallocation.

Contribution of Modality-Relevance Weights. We ablate r_A, r_V, and r_{AV} individually by zeroing the selected component only during base coefficient construction in equation 3, without renormalizing the others. PCME and CAA retain the original relevance vector, leaving C and B un-

Table 3: Comparison of different modality fusion strategies on AVHBench using Qwen2.5-Omni-3B. CAD achieves the highest overall accuracy among all strategies.

Fusion Method	AVHBench		
	Video-Driven Audio Hall.	Audio-Driven Video Hall.	Overall Acc.
Base	74.3	80.2	76.2
Uniform	75.9	**82.2**	78.0
Argmax	78.6	81.8	79.6
CAD	**82.6**	**82.2**	**82.5**

Table 4: Ablation of CAD's modality-relevance weights on AVHBench using Qwen2.5-Omni-7B. Unchecked components are individually zeroed only in equation 3.

Decoding	Weights			AVHBench		Acc.
	r_A	r_V	r_{AV}	Video-Driven Audio Hall.	Audio-Driven Video Hall.	
CAD		✓	✓	80.6	85.8	82.3
CAD	✓		✓	81.4	80.0	80.9
CAD	✓	✓		82.3	84.9	83.2
CAD	✓	✓	✓	**83.2**	**85.9**	**84.1**

Table 5: Ablation of conflict-aware components in CAD on AVHBench using Qwen2.5-Omni-7B. C is potential cross-modal conflict magnitude; B is the conflict actionability score.

Decoding	Conflict Assessment		AVHBench		Acc.
	C	B	Video-Driven Audio Hall.	Audio-Driven Video Hall.	
CAD			79.4	84.7	81.2
CAD	✓		81.5	85.0	82.7
CAD		✓	82.3	85.5	83.3
CAD	✓	✓	**83.2**	**85.9**	**84.1**

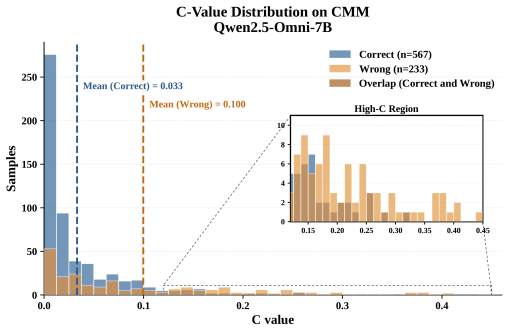

Figure 4: C distributions for correct and wrong Qwen2.5-Omni-7B predictions on CMM.

changed for the same initial branch predictions. Table 4 reports AVHBench results for Qwen2.5-Omni-7B. Removing r_A reduces video-driven audio hallucination accuracy by 2.6 points, while removing r_V yields the largest overall drop of 3.2 points. Removing r_{AV} reduces accuracy in both categories, with a 0.9-point overall drop. The full relevance vector performs best overall, supporting each component's contribution to CAD's coefficient construction in this setting.

Contribution of Conflict-Aware Components. We assess PCME and CAA through four CAD variants: relevance-conditioned decoding without intervention; $B = 1$, yielding $H = C$; $C = 1$ and $\tau = 0$, yielding $H = B$; and full CAD with $H = CB$. As shown in Table 5, using C alone improves overall accuracy from 81.2% to 82.7%, while B alone achieves 83.3%. Combining both reaches 84.1% and improves both hallucination categories, supporting their complementary roles in measuring potential conflict and assessing intervention support.

Effectiveness of Potential Conflict Magnitude. To assess whether C captures unreliable cross-modal fusion, we compare its distributions for correct and wrong Qwen2.5-Omni-7B predictions on CMM. Fig. 4 shows a mean C of 0.100 for errors, over three times the 0.033 for correct predictions. Errors are more concentrated in the high-C region, indicating that stronger disagreement between modality-specific and joint predictions is associated with cross-modal reasoning errors. This difference demonstrates that C characterizes potential cross-modal conflict severity.

6 Conclusion

In this work, we propose Conflict-Aware Decoding (CAD), a training-free framework for mitigating cross-modal hallucinations in Omni-LLMs. CAD assesses potential conflict magnitude and actionability across audio, visual, and joint audio-visual predictions, then selectively reallocates decoding weights based on these signals. Experiments on CMM and AVHBench across four audio-visual models show improvements over training-free baselines for both video-driven audio and audio-driven video hallucinations. Despite these gains, CAD's performance currently depends on manually tuned hyperparameters. Its multi-branch decoding also incurs additional inference overhead relative to native decoding. Future work will explore adaptive parameter selection and optimize the implementation to reduce inference latency and resource consumption.

AI USE STATEMENT

During the preparation of this manuscript, we used AI tools to assist with writing and language polishing, with the aim of improving the clarity and readability of the presentation. The authors have reviewed the AI-assisted text and retrieved information incorporated into the manuscript. The authors take full responsibility for the final content of the paper, including the accuracy of its statements, the interpretation of the reported results, and the conclusions presented.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the method and experimental settings to support reproducibility. Section 4 presents the mathematical formulation and decoding procedure of CAD. Section 5.1 describes the benchmarks, backbone models, hardware, video sampling, and generation settings, while Sections 5.2 and 5.3 provide additional evaluation settings and ablation configurations. Further experimental details are provided in the appendix. Part of the implementation code has been submitted as supplementary material, and the complete implementation will be made publicly available upon acceptance of the paper.

REFERENCES

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, Furong Xu, GuangMing Yao, Jun Zhou, Jingdong Chen, Jianxin Sun, Jiajia Liu, Jianjiang Zhu, Jun Peng, Kaixiang Ji, Kaiyou Song, Kaimeng Ren, Libin Wang, Lixiang Ru, Lele Xie, Longhua Tan, Lyuxin Xue, Lan Wang, Mochen Bai, Ning Gao, Pei Chen, Qingpei Guo, Qinglong Zhang, Qiang Xu, Rui Liu, Ruijie Xiong, Sirui Gao, Tinghao Liu, Taisong Li, Weilong Chai, Xinyu Xiao, Xiaomei Wang, Xiaoxue Chen, Xiao Lu, Xiaoyu Li, Xingning Dong, Xuzheng Yu, Yi Yuan, Yuting Gao, Yunxiao Sun, Yipeng Chen, Yifei Wu, Yongjie Lyu, Ziping Ma, Zipeng Feng, Zhijiang Fang, Zhihao Qiu, Ziyuan Huang, and Zhengyu He. Ming-omni: A unified multimodal model for perception and generation, 2025. URL https://arxiv.org/abs/2506.09344.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Alan Bundy and Lincoln Wallen. *Dempster-Shafer Theory*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1984. ISBN 978-3-642-96868-6. doi: 10.1007/978-3-642-96868-6_52. URL https://doi.org/10.1007/978-3-642-96868-6_52.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24185–24198, June 2024.

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, 2024. URL https://arxiv.org/abs/2406.07476.

Yeongjae Cho, Keonwoo Kim, Taebaek Hwang, and Sungzoon Cho. Do you keep an eye on what i ask? mitigating multimodal hallucination via attention-guided ensemble decoding. In *International Conference on Learning Representations*, volume 2025, pp. 39873–39893,

2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/637b275a6e65924719198188fc939632-Paper-Conference.pdf.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, volume 2024, pp. 54158–54183, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/edc36117f795ca52a0cbf6a7b3882859-Paper-Conference.pdf.

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

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2025a. URL https://arxiv.org/abs/2405.21075.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Yuhang Dai, Meng Zhao, Yi-Fan Zhang, Shaoqi Dong, Yangze Li, Xiong Wang, Haoyu Cao, Di Yin, Long Ma, Xiawu Zheng, Rongrong Ji, Yunsheng Wu, Ran He, Caifeng Shan, and Xing Sun. Vita: Towards open-source interactive omni multimodal llm, 2025b. URL https://arxiv.org/abs/2408.05211.

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18959–18969, 2025.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms, 2026. URL https://arxiv.org/abs/2502.04326.

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

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27036–27046, June 2024.

Chaeyoung Jung, Youngjoon Jang, and Joon Son Chung. Avcd: Mitigating hallucinations in audio-visual large language models through contrastive decoding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 63143–63174. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/5b5e70e7fb3a9060562361b2385346df-Paper-Conference.pdf.

Junho Kim, Hyun J Kim, Yeon J Kim, and Yong M Ro. Code: Contrasting self-generated description to combat hallucination in large multi-modal models. *Advances in Neural Information Processing Systems*, 37:133571–133599, 2024.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, June 2024.

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

in Neural Information Processing Systems, volume 38. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/9b0b18a77421d45d26c3df5612caefe7-Paper-Datasets_and_Benchmarks_Track.pdf.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *Science China Information Sciences*, 68(10):200102, 2025.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12286–12312, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.687. URL https://aclanthology.org/2023.acl-long.687/.

Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, Song Chen, Xu Li, Da Pan, Shusen Zhang, Xin Wu, Zheng Liang, Jun Liu, Tao Zhang, Keer Lu, Yaqi Zhao, Yanjun Shen, Fan Yang, Kaicheng Yu, Tao Lin, Jianhua Xu, Zenan Zhou, and Weipeng Chen. Baichuan-omni technical report, 2024. URL https://arxiv.org/abs/2410.08565.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.342. URL https://aclanthology.org/2024.emnlp-main.342/.

Jing Liu, Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, and Jinhui Tang. Valor: Vision-audio-language omni-perception pretraining model and dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):708–724, 2024.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report, 2026. URL https://arxiv.org/abs/2507.05201.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-salmonn: Speech-enhanced audio-visual large language models, 2024. URL https://arxiv.org/abs/2406.15704.

Kim Sung-Bin, Oh Hyun-Bin, Lee Jung-Mok, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. In *International Conference on Learning Representations*, volume 2025, pp. 24244–24271, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/3cc685788a311fa35d8d41df93e288ca-Paper-Conference.pdf.

Qwen Team. Qwen3.5-omni technical report, 2026. URL https://arxiv.org/abs/2604.15804.

Wenwen Tong, Hewei Guo, Dongchuan Ran, Jiangnan Chen, Jiefan Lu, Kaibin Wang, Keqiang Li, Xiaoxu Zhu, Jiakui Li, Kehan Li, Xueheng Li, Lumin Li, Chenxu Guo, Jiasheng Zhou, Jiandong Chen, Xianye Wu, Jiahao Wang, Silei Wu, Lei Chen, Hanming Deng, Yuxuan Song, Dinghao Zhou, Guiping Zhong, Ken Zheng, Shiyin Kang, and Lewei Lu. Interactiveomni: A unified omni-modal model for audio-visual multi-turn dialogue, 2025. URL https://arxiv.org/abs/2510.13747.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation. In *International Conference on Learning Representations*, volume 2025, pp. 13712–13736, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/24079b91da7257cb78805262996152b8-Paper-Conference.pdf.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15840–15853, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.937. URL https://aclanthology.org/2024.findings-acl.937/.

Zhifei Xie and Changqiao Wu. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities, 2024. URL https://arxiv.org/abs/2410.11190.

Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. Mitigating object hallucination via concentric causal attention. In *Advances in Neural Information Processing Systems*, volume 37, pp. 92012–92035. Curran Associates, Inc., 2024. doi: 10.52202/079017-2920.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025b. URL https://arxiv.org/abs/2509.17765.

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14635–14645, June 2025.

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

YiFan Zhang, Yang Shi, Weichen Yu, Qingsong Wen, Xue Wang, Wenjing Yang, Zhang Zhang, Liang Wang, and Rong Jin. Debiasing multimodal large language models via penalization of language priors. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 4232–4241, New York, NY, USA, 2025. Association for Computing Machinery.

A CAD DECODING PROCEDURE

Algorithm 1 specifies CAD as defined in equation 1–equation 13, including the task-answer projection in equation 8, and assumes that all four contexts AV, A, V, T are available. CAD uses task-answer predictions to assess conflict actionability, then generates a single response by combining full-vocabulary logits with the adjusted coefficients. The Qwen-specific implementation notes below cover tokenization, numerical safeguards, media processing, and scoring.

Algorithm 1 CAD with initial-step conflict assessment

Require: Model θ; audio X_A, video X_V, query X_Q; relevance prompt X_M; tokenizer; maximum output length N; parameters γ, λ, τ; task-answer labels $\mathcal{Y}$ and token sets $\mathcal{T}(y)$.
Ensure: Decoded response string.

Construct AV, A, V, T contexts with the same query and an empty answer prefix.
Evaluate the relevance prompt; obtain (r_A, r_V, r_{AV}) using equation 2.
Prefill each branch and retain its initial next-token logits z_k^1.
Construct the signed coefficients w_k using equation 3.
Form full-vocabulary distributions p_k for $k \in \{A, V, AV\}$.
Set $R_A = r_A + r_{AV}$ and $R_V = r_V + r_{AV}$.
Compute g_{rel} and p_{mix} using equation 5 and equation 6.
Set $D_{A,V} = \mathrm{JSD}_n(p_A, p_V)$ and $D_F = \mathrm{JSD}_n(p_{AV}, p_{\mathrm{mix}})$; compute C using equation 7.
Project p_A, p_V, p_{AV} onto $\mathcal{Y}$ using equation 8 to obtain π_A, π_V, π_{AV}.
Set $y_k = \arg\max_{y \in \mathcal{Y}} \pi_k(y)$ and $m_k = \pi_k^{(1)} - \pi_k^{(2)}$; compute $q_A = R_A m_A$, $q_V = R_V m_V$.
if $y_A = y_V = y_{AV}$ **then**
    $B \leftarrow 0$.
else if $y_A = y_V$ **then**
    $B \leftarrow \min(q_A, q_V)$.
else if $y_{AV} = y_A$ **then**
    $B \leftarrow [q_V - q_A]_+$.
else if $y_{AV} = y_V$ **then**
    $B \leftarrow [q_A - q_V]_+$.
else
    $B \leftarrow \max(q_A, q_V)$.
end if
$H \leftarrow CB$; $\widetilde{w} \leftarrow w$.
if $C \geq \tau$, $C > \epsilon$, $H > \epsilon$, $\lambda > 0$, and the answer gate reports valid distributions (numerical safeguard) **then**
    $\delta \leftarrow \lambda H [w_{AV}]_+$.
    if $\delta > \epsilon$ **then**
        Compute (ρ_A, ρ_V) using equation 11, using equal shares when $[w_A]_+ + [w_V]_+ \leq \epsilon$.
        Apply equation 12 to obtain $\widetilde{w}$.
    end if
end if
for $t = 1, \ldots, N$ **do**
    Select $\widehat{y}_t = \arg\max_{v \in \mathcal{V}} \sum_k \widetilde{w}_k z_k^t(v)$.
    Append $\widehat{y}_t$; stop if it is the end-of-sequence token or $t = N$.
    Append this same token to every branch prefix and compute the next branch logits.
end for
Decode the selected tokens, skipping special tokens, and return the string.

Relevance scores are obtained from a dedicated prompt. Before selecting the first response token, the answer branches are evaluated on their respective contexts with an empty generated-answer prefix. PCME computes C from the full-vocabulary next-token distributions of A, V, AV. CAA computes B from the task-answer predictions and margins derived from the projected distributions in equation 8, together with the effective relevance scores. Relevance, C, B, and the resulting coefficients $\widetilde{w}$ remain fixed throughout generation. At each subsequent step, every branch retains its own input context and updates its logits using the same previously selected response tokens.

The branch coefficients are signed logit multipliers. The update in equation 12 preserves their sum because $\rho_A + \rho_V = 1$, and leaves w_T unchanged. The transfer $\delta = \lambda CB[w_{AV}]_+$ is not capped at $[w_{AV}]_+$. For $w_{AV} > 0$, an active update gives $\widetilde{w}_{AV} = w_{AV}(1 - \lambda CB)$, which is negative when $\lambda CB > 1$. When $[w_A]_+ + [w_V]_+ > \epsilon$, the transfer is allocated between the audio and video branches in proportion to these positive parts; otherwise, $\rho_A = \rho_V = 1/2$, so each branch receives $\delta/2$. Thus, even with C, B, and λ fixed, changes to w_{AV} can change the transfer amount, and changes to w_A, w_V can change its allocation.

B Implementation and numerical controls

Inputs and branch construction. The implementation details below concern the Qwen adapter. The four answer branches retain the same query and system message. AV receives video and audio, A receives audio, V receives video, and T receives only text. For the relevance head, the user text is the question preceded by `Question:` and followed by the prompt: "To answer this question, which modality is needed (audio, video, or both):". The binary CMM wrapper appends "Answer only 'Yes' or 'No'. Do not include any explanation." to the question; the AVHBench wrapper does so by default for non-captioning tasks, with an option to disable that suffix.

Parameters and configuration precedence. Table 6 summarizes the CAD parameters and numerical controls. The listed hyperparameter values define a reference configuration. Hyperparameters for the reported experiments were selected on a separate validation set before evaluation. The numerical constant $\epsilon = 10^{-8}$ is specified in Section 4.2 and used by the Qwen implementation. The task-answer space $\mathcal{Y}$ and token sets $\mathcal{T}(y)$ are described alongside equation 8, which defines the projected distributions used by CAA. Both Qwen evaluation wrappers pass γ and λ explicitly to the decoder. The decoder reads τ, N, and video-sampling settings from the parsed arguments. An explicit value of λ takes precedence over the value in the argument object.

Table 6: CAD parameters and numerical controls in the Qwen implementation, with the reference configuration $\gamma = 4.1$, $\lambda = 7.9$, and $\tau = 0.05$.

Control	Meaning and implementation role
γ	Relevance-conditioned contrast strength in equation 3; changes the base coefficients and hence the redistribution shares.
λ	Scales the transfer $\delta = \lambda CB[w_{AV}]_+$; does not change the initial branch distributions.
τ	Lower bound on raw C for intervention, not a threshold on CB or label probability.
ϵ	10^{-8} in numerical probability, divergence, and intervention guards.
N	Maximum number of generated response tokens.

Numerical operations. The probability helper subtracts the maximum logit before exponentiation and adds ϵ to its denominator. Distribution-normalization helpers clip negative inputs to zero and normalize their sum; they return a uniform vector when a nonempty vector has total mass at most ϵ. The JSD helper normalizes its inputs, adds ϵ inside logarithms, divides by $\log 2$, and clips the result to $[0, 1]$. PCME and the reliability helper normalize their input relevance vector before computing effective relevance. In the Table 4 ablations, they receive the original vector; only equation 3 receives the separately masked, unnormalized copy. Thus, C and B remain unchanged for the same initial branch predictions, whereas the modified base coefficients can change the transfer amount $\delta = \lambda CB[w_{AV}]_+$ when intervention is active and the redistribution shares (ρ_A, ρ_V) in equation 11. These changes affect the adjusted coefficients in equation 12. All four branches are retained, and the hyperparameters are fixed across variants.

The reliability helper checks that its input probability vectors are nonempty, finite, one-dimensional, have positive mass after clipping, and have the same number of entries. Invalid inputs suppress answer-gated intervention. It clips reliabilities and the returned gate to $[0, 1]$. The intervention path also treats C, H, and δ no larger than ϵ as inactive. For redistribution, a positive-coefficient sum no larger than ϵ invokes the equal-share fallback when both unimodal branches exist. These are numerical guards around the equations, rather than additional learned parameters.

Model calls and caching. The Qwen wrappers load the model in bfloat16 and obtain logits from its text-generating component. Branch probabilities for the diagnostics are computed after converting logits to floating-point values. For a one-token output in the branch-decoding modes, the answer-branch prefill calls disable the key/value cache. With a longer output, the decoder stores a separate cache per branch and feeds each branch the same selected next token. This cache setting concerns the answer branches; the relevance-head call is separate. All available answer branches are evaluated, including a branch whose decoding coefficient happens to be zero. No measured latency or memory reduction follows solely from this control flow.

C Answer Scoring and Decoding

Task-answer labels and token sets. The task defines the shared answer space $\mathcal{Y}$: yes/no labels for binary tasks and the valid option labels for each multiple-choice question. The token-set construction described below follows the Qwen adapter.

For binary labels, the adapter considers lowercase and initial-capital forms, each with an optional single leading space and an optional final period. For multiple-choice labels, it considers the uppercase letter alone, followed by a period, or enclosed in parentheses, each with or without a single leading space.

We encode these candidate forms using the model's tokenizer and retain only those encoded as a single token. Token IDs are deduplicated within each label to construct $\mathcal{T}(y)$, so each token's probability contributes only once to the score for that label.

Initial distributions and conflict assessment. Before generating the first response token, we convert the initial logits of branches $k \in \{A, V, AV\}$ into full-vocabulary probability distributions p_k. PCME uses these distributions to compute the conflict magnitude C. For CAA, we sum the probabilities of the token IDs in $\mathcal{T}(y)$ for each label y and normalize across $\mathcal{Y}$ according to equation 8, obtaining the task-answer distributions π_k.

Each π_k yields a predicted label $y_k = \arg\max_{y \in \mathcal{Y}} \pi_k(y)$ and a probability margin $m_k = \pi_k^{(1)} - \pi_k^{(2)}$, where $\pi_k^{(1)}$ and $\pi_k^{(2)}$ are its two largest probabilities. The reliability proxies $q_A = R_A m_A$ and $q_V = R_V m_V$, together with the relations among the branch predictions, determine the conflict actionability score B in equation 10. The resulting signal $H = CB$ is used for coefficient adjustment through equation 11 and equation 12.

Response generation and evaluation. The adjusted coefficients remain fixed throughout generation. At each decoding step, CAD combines the branch logits using these coefficients and selects the next token by the full-vocabulary argmax in equation 13. The tokenizer decodes the generated token sequence into text, omitting special tokens. The evaluation parser then extracts the predicted task label from this text for comparison with the ground-truth label.