Modality-Grounded Contrastive Decoding for Cross-Modal Hallucination Mitigation

Yuyao Ge[♣] **Shenghua Liu**[♣*] **Yiwei Wang**[♢] **Baolong Bi**[♣]
Lingrui Mei[♣] **Jiayu Yao**[♣] **Xueqi Cheng**[♣]

[♣]Institute of Computing Technology [♢]University of California, Merced
{geyuyao24z, liushenghua}@ict.ac.cn

Abstract

Multimodal large language models suffer from cross-modal hallucinations, where signals from one modality bias predictions about another. To quantify this, we decompose fused logits into per-modality contributions and define a *Cross-Modal Interference* (CMI) ratio that measures the relative dominance of the non-target modality in the fused prediction margin. CMI provides a per-sample diagnostic of cross-modal interference: high-CMI samples exhibit error rates 19–39.5% higher than low-CMI samples. Further analysis shows that the target modality's independent judgment carries critical corrective information, but exploiting it effectively requires continuous soft correction; hard override degrades accuracy by up to 9.6%. We propose *Modality-Grounded Contrastive Decoding* (MGCD), a training-free framework that softly anchors the fused prediction toward the target modality's own judgment when the former overshoots, while preserving the original prediction otherwise. MGCD introduces only a single additional hyperparameter with an analytically characterized effective range. MGCD generalizes across both end-to-end and modular architectures. Systematic comparison of six grounding strategies reveals that anchoring to the target modality margin yields the largest and most consistent improvement, with a clear gap over strategies that lack modality-specific directional information. Across architecturally diverse models and multiple benchmarks, MGCD recovers 50–70% of the oracle ceiling on fixable errors and consistently improves over all baselines on aggregate metrics, with gains of up to 11.4% over standard decoding. Our results demonstrate that the form of correction, continuous anchoring rather than binary override, is at least as important as the signal itself for inference-time hallucination mitigation, a design axis that existing contrastive decoding methods have not explored.

1 Introduction

Multimodal large language models (MLLMs) that jointly process video, audio, and text have enabled impressive audio-visual reasoning capabilities [Zhang et al., 2023a, Chu et al., 2024, Geng et al., 2025]. However, these models exhibit a persistent failure mode: *cross-modal hallucinations* [Sung-Bin et al., 2025, Leng et al., 2024a], where one modality's signal inappropriately influences predictions about another. For instance, a model may see a dog in the video frame and hallucinate hearing it bark, even when the audio contains only ambient noise.

Unlike language-prior hallucinations that fabricate content from parametric memory [Li et al., 2023a, Guan et al., 2024], cross-modal hallucinations arise from *inter-modal interference* during the fusion of modality-specific representations. Contrastive decoding [Li et al., 2023b, Leng et al., 2024b] is

*Corresponding author.

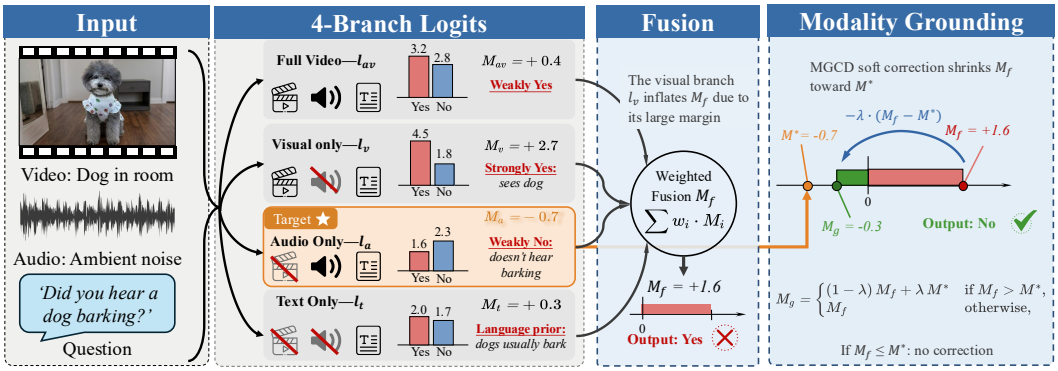


Figure 1: Overview of MGCD. The model extracts logits from four branches with different modality masks, computes per-branch confidence margins, and fuses them into a single decision margin. MGCD then applies a soft correction that shrinks the fused margin toward the target modality's own margin when the former overshoots, yielding the grounded margin used for the final prediction.

the predominant training-free mitigation approach, suppressing hallucinations by contrasting the output distribution of the full input against a degraded input (e.g., with a modality masked). Recent extensions to multi-modal settings include AVCD [Jung et al., 2025] and MAD [Chung et al., 2026], which introduce modality-aware branch weighting.

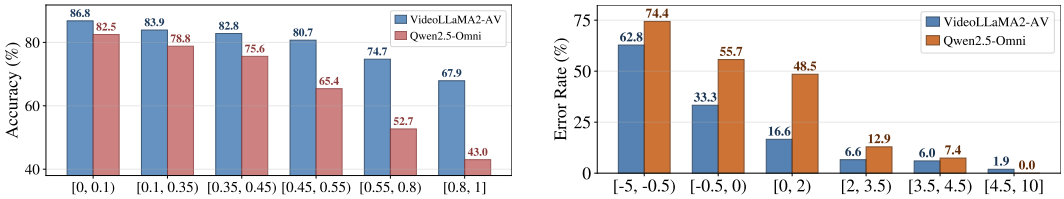
Existing contrastive decoding methods focus on *how to fuse* multi-branch logits, but the *margin difference across branches* provides an untapped corrective signal encoding whether the target modality agrees with the fused prediction. Applying this signal naively is disastrous: binary override based on the target modality margin degrades accuracy by up to 9.6%, because many flagged samples are actually correct.

We decompose fused logits into per-modality contributions and define *Cross-Modal Interference* (CMI), revealing that high-interference samples concentrate prediction errors (Section 3.1). The target modality's branch margin further provides per-sample correction direction: strong disagreement with the fused prediction correlates with error rates exceeding 55%. Unlike existing methods that focus on how to weight multi-branch outputs, we shift the operation from logit space to margin space and introduce per-sample directional correction guided by the target modality's independent judgment. We propose **Modality-Grounded Contrastive Decoding** (MGCD), illustrated in Figure 1, which softly anchors the fused margin toward the target modality's judgment when the former overshoots, while preserving the original prediction otherwise. Our contributions are threefold:

- We define CMI via logit decomposition and show that high-CMI samples exhibit 19–39.5% higher error rates; the target modality margin further provides per-sample correction direction.
- We formalize the correction as a signal-noise interpolation problem and prove that soft anchoring improves the signal-to-noise ratio (SNR), with an analytically characterized effective range for the anchoring strength.
- Through systematic comparison of six grounding strategies and two correction mechanisms, we identify three design axes, namely anchor signal, correction form, and operational space, and instantiate them as MGCD: a single-equation framework that recovers 50–70% of the oracle ceiling across both models with one additional hyperparameter.

2 Related Work

Hallucination mitigation at inference time. Contrastive decoding [Li et al., 2023b] suppresses degenerate patterns by contrasting expert and amateur distributions; DeCK [Bi et al., 2025] extends this to knowledge editing. In the vision-language domain, VCD [Leng et al., 2024b] contrasts outputs from original and distorted visual inputs, and ICD [Wang et al., 2024] contrasts standard and instruction-disturbed prompts to subtract hallucinated concepts. DoLa [Chuang et al., 2023] contrasts later transformer layers against earlier ones, while ITI [Li et al., 2023c] shifts activations along



(a) Accuracy by CMI bin.

(b) Error rate by target margin M^*.

Figure 2: Cross-modal interference analysis (V→A + pred=Yes). **(a)** CMI $\in [0, 1]$: higher values indicate greater non-target dominance. **(b)** M^*: negative values indicate the target modality opposes the fused prediction. Both metrics show clear declining trends on both models.

probed truthful directions. For multimodal settings, OPERA [Huang et al., 2024] penalizes logits tied to attention over-trust, PAI [Liu et al., 2024] amplifies image-token attention while subtracting text-only logits, HALC [Chen et al., 2024] combines auto-focal grounding with beam search, and M3ID [Favero et al., 2024] maximizes mutual information between generated tokens and visual input. For audio-visual models, AVCD [Jung et al., 2025] introduces multi-branch contrastive objectives with attention-guided masking over CLIP [Radford et al., 2021] and Whisper [Radford et al., 2023], and MAD [Chung et al., 2026] adds dynamic modality weighting via self-assessed relevance. These methods focus on how to combine or contrast multi-branch outputs; none exploits the *margin differences across branches* as a per-sample correction signal.

Modality imbalance in multimodal fusion. Multimodal networks tend to rely on the easiest modality [Wu et al., 2022]; OGM-GE [Peng et al., 2022], MMPareto [Wei and Hu, 2024], QMF [Zhang et al., 2023b], PDF [Cao et al., 2024], and EAU [Gao et al., 2024] address this via gradient modulation, Pareto objectives, quality-based reweighting, per-modality calibration, and aleatoric uncertainty estimation, respectively. These approaches require fine-tuning or auxiliary heads, whereas MGCD operates at decoding time on a frozen model. The soft anchoring relates to James–Stein shrinkage [Stein, 1956, James et al., 1961]; the key departures are the one-sided gate and the sample-dependent target margin anchor. More broadly, MGCD is, to our knowledge, the first to shift contrastive decoding to margin space, where the signal-to-noise structure admits formal analysis (Proposition 1) and branch margins provide per-sample adaptive correction.

3 Method

3.1 Cross-Modal Interference Analysis

Notation. Let f denote the MLLM and $\varnothing$ a null (masked) modality input. Given input $x = (x_v, x_a, x_t)$, where x_v, x_a, x_t denote the visual, audio, and text components, a multi-branch contrastive decoder produces four logit vectors l_{av}, l_v, l_a, l_t by selectively masking modalities. The *branch margin* for branch $i \in \{av, v, a, t\}$ over two candidate answers y_1, y_2, where y_1 is the higher-scoring option under the full branch l_{av} (e.g., Yes in a Yes/No task), is:

$$M_i = l_i(y_1) - l_i(y_2). \tag{1}$$

The *target modality margin* M^* is defined as M_a for V→A tasks and M_v for A→V tasks, corresponding to the questioned modality. Branch margins are fused as $M_f = \sum_{i \in \{av,v,a,t\}} w_i(x)\, M_i$. The fusion weights w_i can follow various strategies; in the default configuration, we use relevance-weighted fusion $w_i(x) = \gamma \cdot p_i(x)$, where $p_i(x)$ is obtained by querying the model for modality relevance (Appendix B), γ is the contrastive strength, and $w_t = -1$. In the base form γ is fixed across branches; Algorithm 1 uses an entropy-adaptive variant γ_i described in Practical Refinements (Section 3.2).

Table 1: Grounding strategy comparison on AVHBench with QWEN2.5-OMNI (%). Six strategies are compared, differing in the anchor signal used to correct M_f. Δ is relative to No grounding. Anchoring to the target modality margin M^* yields the largest gain; the +0.5% gap over fused-only isolates the value of modality-specific directional information.

Strategy	Description	Acc.	Δ
No grounding (λ=0)	$M_g = M_f$	82.4	—
Interfere-branch	$M_g = M_f - \lambda \max(0, M_f - M_{\text{non-tgt}})$	81.9	−0.5
Random	$M_g = M_f - \lambda \max(0, M_f - R)$, $R \sim \mathcal{N}(0,1)$	82.2	−0.2
Fixed	$M_g = M_f - \lambda \max(0, M_f - c)$, c=0	82.9	+0.5
Fused-only	$M_g = (1-\lambda) M_f$	82.9	+0.5
Target (MGCD)	$M_g = M_f - \lambda \max(0, M_f - M^*)$	**83.4**	**+1.0**

Table 2: Continuous grounding compared with binary override on AVHBench (%). The baseline is MAD [Chung et al., 2026], the strongest existing method. All four binary thresholds degrade accuracy on both models, while MGCD is the only strategy that improves. Note that MGCD's margin-level fusion *without* grounding (λ=0) already achieves 82.4% on QWEN2.5-OMNI, 1.8% above MAD's logit-level fusion; the grounding step adds a further 1.0%.

Strategy	VIDEOLLAMA2-AV Acc.	Δ	QWEN2.5-OMNI Acc.	Δ
Baseline	79.6	—	80.6	—
$M^* < 0 \rightarrow$ flip	71.8	−7.8	71.0	−9.6
$M^* < -0.5 \rightarrow$ flip	73.8	−5.8	71.0	−9.6
$M^* < -1.0 \rightarrow$ flip	75.2	−4.4	71.2	−9.4
$M^* < -2.0 \rightarrow$ flip	76.0	−3.6	72.8	−7.8
MGCD	**82.2**	**+2.6**	**83.4**	**+2.8**

To quantify the role of each modality in the fused prediction, we decompose l_{av} into per-modality contributions:

$$\begin{aligned} l_{av}(y) = l_t(y) + \underbrace{(l_v(y) - l_t(y))}_{C_v} + \underbrace{(l_a(y) - l_t(y))}_{C_a} \\ + \underbrace{(l_{av}(y) - l_v(y) - l_a(y) + l_t(y))}_{C_I}, \end{aligned} \tag{2}$$

This additive attribution follows the inclusion-exclusion expansion underlying Shapley-based attribution [Lundberg and Lee, 2017] for a two-player cooperative game between the visual and audio modalities, with C_I capturing their interaction (see Appendix A for the precise characterization and margin-level expansion). We define the Cross-Modal Interference (CMI) ratio from the margin-level counterparts $C_i^M = C_i(y_1) - C_i(y_2)$. For a V→A task the target modality is audio, so $C^M_{\text{target}} = C^M_a$ and $C^M_{\text{non-target}} = C^M_v$ (the interaction term C^M_I is excluded to isolate single-modality contributions):

$$\text{CMI} = \frac{|C^M_{\text{non-target}}|}{|C^M_{\text{non-target}}| + |C^M_{\text{target}}|} \in [0, 1], \tag{3}$$

with CMI defined as 0 when both terms vanish. Figure 2(a) stratifies accuracy by CMI on the V→A + pred=Yes subset, the primary cross-modal hallucination scenario. Accuracy shows a clear declining trend on both models, with a gap of 19.0% for VIDEOLLAMA2-AV and 39.5% for QWEN2.5-OMNI.

CMI identifies interference at the population level; correction requires a per-sample signal. Figure 2(b) shows that M^* serves this role: error rate exceeds 55% when $M^* < -0.5$ and drops to single digits as M^* grows; Table 10 (Appendix C) provides per-bucket statistics for VIDEOLLAMA2-AV. M^* thus predicts errors and indicates correction direction. *How* the signal is used matters equally: binary override causes catastrophic degradation (Table 2), while continuous grounding yields gains.

3.2 Modality-Grounded Contrastive Decoding

We now formalize *why* continuous grounding works. The CMI analysis suggests that M_f is contaminated by non-target modality noise while M^* is relatively clean. We model this with a tractable

signal framework. Assume the true label is $y^* \in \{+1, -1\}$ and the two margins are generated as:

$$M_f = y^*\,\delta_f + \epsilon_f, \qquad M^* = y^*\,\delta^* + \epsilon^*, \tag{4}$$

where $\delta_f, \delta^* > 0$ are signal strengths, ϵ_f, ϵ^* are independent zero-mean noise with standard deviations σ_f and σ^*, and $\sigma_f > \sigma^*$, consistent with the CMI analysis showing higher error rates when non-target contributions dominate (see Appendix A for the causal justification). Proposition 1 below requires only zero-mean noise with finite variance; the Gaussian assumption is not needed for the SNR result. In practice, branch margins arise from nonlinear transformations of shared representations, so the zero-mean condition is approximate; the empirical consistency between the predicted effective range and the observed plateau in Figure 5 (Appendix D) supports the model's practical relevance. In practice M_f includes M^* as a weighted component, introducing positive correlation $\rho > 0$. Corollary 3 (Appendix E) shows that the SNR improvement holds whenever $r > \rho\, s$, with a narrower but positive effective interval $\lambda_{\max}(\rho)$; the empirical gains confirm this across both architectures.

Modality-grounded margin. We define the *modality-grounded margin*:

$$M_g = \begin{cases} (1-\lambda)\,M_f + \lambda\,M^* & \text{if } M_f > M^*, \\ M_f & \text{otherwise,} \end{cases} \tag{5}$$

where $\lambda > 0$ controls the anchoring strength and the one-sided gate restricts correction to the overshoot subset, where the error rate is substantially higher than on the complement. For $\lambda \in (0, 1)$ the correction is a convex combination of M_f and M^*; $\lambda > 1$ permits over-correction beyond M^*, which can be beneficial when the target margin itself underestimates the true signal. Equivalently, in penalty form:

$$M_g = M_f - \lambda \cdot \max(0,\ M_f - M^*). \tag{6}$$

The final prediction is $\hat{y} = y_1$ if $M_g > 0$, else y_2. λ uses $\gamma/3$ as its starting point, with a performance plateau around this value (Section D; full settings in Table 11, Appendix F).

SNR analysis. The SNR of M_f under the signal model is δ_f/σ_f. The interpolation branch of Eq. 5, $(1-\lambda)M_f + \lambda M^*$, has SNR:

$$\text{SNR}(M_g) = \frac{(1-\lambda)\,\delta_f + \lambda\,\delta^*}{\sqrt{(1-\lambda)^2\,\sigma_f^2 + \lambda^2\,\sigma^{*2}}}. \tag{7}$$

Let $r = \delta^*/\delta_f$ (signal ratio) and $s = \sigma^*/\sigma_f < 1$ (noise ratio). $\text{SNR}(M_g) > \text{SNR}(M_f) = \delta_f/\sigma_f$ reduces to:

$$2(1-\lambda)\,r + \lambda\,(r^2 - s^2) > 0. \tag{8}$$

Proposition 1 (Effective anchoring interval)**.** *If $r > 0$ (target modality carries signal), there exists $\lambda_{\max} > 0$ such that the grounded margin improves the SNR for all $\lambda \in (0, \lambda_{\max})$. Specifically:* (i) *if $r \geq s$, then $\lambda_{\max} = 1$, i.e., the full interpolation range is effective;* (ii) *if $r < s$, then $\lambda_{\max} = 2r\,/\,(2r + s^2 - r^2)$.*

Proof sketch (full proof in Appendix E). Eq. 8 rewrites as $\lambda(r^2 - s^2) + 2(1-\lambda)r > 0$. If $r \geq s$, both terms are non-negative for all $\lambda \in (0, 1)$, giving $\lambda_{\max} = 1$. If $r < s$, the inequality fails when $\lambda > 2r/(2r + s^2 - r^2)$.

The linear combination is thus a natural choice; more complex forms such as sigmoid gating add parameters without benefit.

Remark 2 (One-sided gating preserves the improvement)**.** *Proposition 1 analyzes the unconditional interpolation; the one-sided gate in Eq. 5 partitions samples into two disjoint subsets: $\mathcal{O}^c = \{M_f \leq M^*\}$, where $M_g = M_f$ and accuracy is unchanged, and $\mathcal{O} = \{M_f > M^*\}$, where the interpolation is applied. Overall accuracy improves if and only if correction improves accuracy on $\mathcal{O}$.*

On $\mathcal{O}$, two cases arise under the signal model: (i) *When the prediction is correct ($y^*{=}{+}1$), both M_f and M^* are typically positive, and the condition $M_f > M^*$ implies fused overconfidence; for $\lambda < 1$ the corrected margin M_g remains positive, so the prediction is preserved.* (ii) *When the prediction is incorrect ($y^*{=}{-}1$), $M_f > 0$ while $M^* < 0$ with high probability (Table 10), so $M_f > M^*$ is easily satisfied and the correction pushes M_g toward sign reversal.*

The gate thus concentrates correction on the error-prone subset while leaving correct predictions largely intact. Table 12 (Appendix I) validates this: correction precision is 71.8–77.8%, showing that the majority of margin flips are beneficial (see Appendix A for a discussion of the approximation).

Algorithm 1 MGCD: Modality-Grounded Contrastive Decoding

Notation: f: MLLM; $\varnothing$: masked modality; Prompt$(\cdot)$: relevance query; q_i: output distribution of branch i; $H(\cdot)$: entropy; K: candidate token count; d: task direction; ϕ: directional flag (restricts grounding to V→A when true)

Require: $x = (x_v, x_a, x_t)$, $d \in \{\text{V}\to\text{A}, \text{A}\to\text{V}\}$, γ, λ, directional flag ϕ
Ensure: Prediction $\hat{y}$

Extract: $l_{av} \leftarrow f(x_v, x_a, x_t)$ ▷ Full audio-visual branch
Extract: $l_v \leftarrow f(x_v, \varnothing, x_t)$ ▷ Visual-only branch
Extract: $l_a \leftarrow f(\varnothing, x_a, x_t)$ ▷ Audio-only branch
Extract: $l_t \leftarrow f(\varnothing, \varnothing, x_t)$ ▷ Text-only (language prior) branch
Query: $\{p_{av}, p_v, p_a\} \leftarrow \text{Softmax}\big(\text{Prompt}(x)\big)$ ▷ Self-assessed relevance
Identify: $y_1, y_2 \leftarrow \text{Top-2}(l_{av})$ ▷ Top-two candidates from fused logits
for each branch $i \in \{av, v, a, t\}$ **do**
  Margin: $M_i \leftarrow l_i(y_1) - l_i(y_2)$ ▷ Branch confidence margin
end for
for each branch $i \in \{av, v, a\}$ **do**
  Adapt: $\gamma_i \leftarrow \gamma \cdot \big(1 - H(q_i)/\log K\big)$ ▷ Entropy-adaptive strength
  Weight: $w_i \leftarrow \gamma_i \cdot p_i(x)$ ▷ Per-branch fusion weight
end for
Contrast: $w_t \leftarrow -1$ ▷ Contrastive text-branch weight
Fuse: $M_f \leftarrow \sum_i w_i \cdot M_i$ ▷ Weighted margin fusion
Select: $M^* \leftarrow M_a$ if d=V→A, else M_v ▷ Target modality margin
if $\phi =$ true **and** $d \neq$ V→A **then**
  $M_g \leftarrow M_f$ ▷ Skip grounding for non-dominant direction
else
  Ground: $M_g \leftarrow M_f - \lambda \cdot \max(0,\ M_f - M^*)$ ▷ Soft one-sided anchoring
end if
Decide: return $\hat{y} = y_1$ if $M_g > 0$, else y_2 ▷ Margin-based decision

Extension to multiple-choice questions. The formulation above targets binary (Yes/No) tasks. For multiple-choice questions with options $\{y^{(1)}, y^{(2)}, \ldots\}$, we first identify the top-two options $y^{(1)}, y^{(2)}$ from the fused logits l_{av}, then compute per-branch margins using the same option pair:

$$M_i = l_i(y^{(1)}) - l_i(y^{(2)}). \tag{9}$$

Using a shared option pair across all branches ensures that the margins are semantically comparable and that $M_f = \sum_i w_i M_i$ remains well-defined. The computation of M_g in Eq. 5 remains unchanged, and the final prediction is $y^{(1)}$ if $M_g > 0$, else $y^{(2)}$; when the number of options is two, this reduces to the binary case. Algorithm 1 summarizes the complete MGCD pipeline.

Practical refinements. Two optional refinements are used in the main evaluation. *Directional activation* introduces a flag $\phi \in \{\text{true}, \text{false}\}$; when $\phi =$ true, the grounding step in Eq. 5 is restricted to V→A, the dominant hallucination direction, since A→V tasks have fewer fixable errors. The remaining pipeline stages still apply to both directions, so any A→V gains under $\phi =$ true reflect multi-branch contrastive decoding rather than the grounding step. *Entropy-adaptive* γ replaces the fixed contrastive strength with a per-branch value $\gamma_i = \gamma \cdot (1 - H(q_i)/\log K)$, where q_i is branch i's output distribution over K candidate tokens and $H(q_i)$ its entropy. Here K=2 for binary and open-ended tasks, corresponding to the top-two candidate tokens from l_{av}, and K=4 for four-option multiple-choice. Low-entropy branches receive higher weight, upweighting confident modality judgments; this provides a small complementary improvement.

3.3 Design Space Analysis

Table 1 compares six strategies differing in the anchor signal. Anchoring to the non-target modality (Interfere-branch) degrades accuracy by 0.5%, confirming that grounding to the wrong modality amplifies interference. Fixed-threshold and fused-only anchoring both improve by +0.5%, but these position-independent corrections capture only half the gain of MGCD at +1.0%; the remaining

Table 3: Main results on AVHBench and CMM (%). **Bold** indicates the best result and underline the second best in each column.

Method	CMM				AVHBench		
	Vis.	Aud.	Lang.	Avg.	V→A	A→V	Avg.
QWEN2.5-OMNI-7B							
Vanilla	64.5	72.3	81.3	72.7	73.0	80.7	75.5
+ VCD_{Ext}	$62.5_{\blacktriangledown 2.0}$	$71.3_{\blacktriangledown 1.0}$	**84.5**$_{\blacktriangle 3.2}$	$72.8_{\blacktriangle 0.1}$	$70.3_{\blacktriangledown 2.7}$	$77.1_{\blacktriangledown 3.6}$	$72.5_{\blacktriangledown 3.0}$
+ AVCD	$66.3_{\blacktriangle 1.8}$	$72.8_{\blacktriangle 0.5}$	$81.0_{\blacktriangledown 0.3}$	$73.4_{\blacktriangle 0.7}$	$75.8_{\blacktriangle 2.8}$	$79.7_{\blacktriangledown 1.0}$	$77.1_{\blacktriangle 1.6}$
+ MAD	$76.8_{\blacktriangle 12.3}$	$84.3_{\blacktriangle 12.0}$	$83.3_{\blacktriangle 2.0}$	$81.5_{\blacktriangle 8.8}$	$78.7_{\blacktriangle 5.7}$	$84.4_{\blacktriangle 3.7}$	$80.6_{\blacktriangle 5.1}$
+ MGCD	**83.5**$_{\blacktriangle 19.0}$	**85.2**$_{\blacktriangle 12.9}$	$83.6_{\blacktriangle 2.3}$	**84.1**$_{\blacktriangle 11.4}$	**82.5**$_{\blacktriangle 9.5}$	**85.2**$_{\blacktriangle 4.5}$	**83.4**$_{\blacktriangle 7.9}$
VIDEOLLAMA2-AV-7B							
Vanilla	71.8	80.0	68.8	73.5	75.7	79.0	76.8
+ VCD_{Ext}	$71.3_{\blacktriangledown 0.5}$	$83.3_{\blacktriangle 3.3}$	**74.8**$_{\blacktriangle 6.0}$	$76.5_{\blacktriangle 3.0}$	$66.0_{\blacktriangledown 9.7}$	$74.8_{\blacktriangledown 4.2}$	$68.9_{\blacktriangledown 7.9}$
+ AVCD	$71.8_{\blacktriangle 0.0}$	$84.0_{\blacktriangle 4.0}$	$71.5_{\blacktriangle 2.7}$	$75.8_{\blacktriangle 2.3}$	$78.3_{\blacktriangle 2.6}$	$80.3_{\blacktriangle 1.3}$	$79.0_{\blacktriangle 2.2}$
+ MAD	$88.8_{\blacktriangle 17.0}$	$84.3_{\blacktriangle 4.3}$	$71.0_{\blacktriangle 2.2}$	$81.4_{\blacktriangle 7.9}$	$80.3_{\blacktriangle 4.6}$	$78.3_{\blacktriangledown 0.7}$	$79.6_{\blacktriangle 2.8}$
+ MGCD	**90.7**$_{\blacktriangle 18.9}$	**84.8**$_{\blacktriangle 4.8}$	$73.9_{\blacktriangle 5.1}$	**83.1**$_{\blacktriangle 9.6}$	**82.9**$_{\blacktriangle 7.2}$	**80.7**$_{\blacktriangle 1.7}$	**82.2**$_{\blacktriangle 5.4}$

+0.5% comes from target-modality directional information. Table 2 tests the correction form: binary override at four thresholds all cause degradation on both models, with losses up to 9.6%. Even the most conservative threshold ($M^* < -2.0$) degrades by 3.6–7.8%, because a substantial fraction of flagged samples are actually correct. MGCD's soft convex combination flips only the most marginal samples while gently correcting the rest, yielding a 6.2–10.6% gap over the best binary variant.

4 Experiments

4.1 Setup

Models. We evaluate on **QWEN2.5-OMNI-7B** [Xu et al., 2025], an end-to-end model with a single transformer backbone, and **VIDEOLLAMA2-AV-7B** [Cheng et al., 2024], a modular model with separate visual and audio encoders. Hyperparameter settings appear in Table 11 (Appendix F).

Datasets. **AVHBench** [Sung-Bin et al., 2025]: 3,419 Yes/No cross-modal hallucination questions (2,287 V→A, 1,132 A→V). **CMM** [Leng et al., 2024a]: 1,200 Yes/No questions across three cross-modal domains (Visual, Audio, Language). **OmniBench** [Li et al., 2024], **WorldSense** [Benchekroun et al., 2023], and **MUSIC-AVQA** [Li et al., 2022] validate generalization to multiple-choice and open-ended tasks. Sub-category details are in Appendix G.

Baselines. **Vanilla**: standard decoding; **VCD_{Ext}** [Leng et al., 2024b]: visual contrastive decoding extended to audio-visual input; **AVCD** [Jung et al., 2025]: multi-branch contrastive objectives with attention-guided masking; **MAD** [Chung et al., 2026]: dynamic branch weighting via self-assessed modality relevance. All baselines use public code and official weights under identical protocols; baseline hyperparameters follow their original papers without additional tuning.

4.2 Main Results

Table 3 presents the main results on AVHBench and CMM. On AVHBench, V→A improves by +3.8% and A→V by +0.8%, consistent with V→A being the dominant hallucination direction.

On VIDEOLLAMA2-AV, MGCD improves by 2.6% on AVHBench and 1.7% on CMM Avg. over MAD, with gains on both hallucination directions. On QWEN2.5-OMNI, MGCD achieves the highest accuracy on six of the seven columns. The largest gain appears on CMM

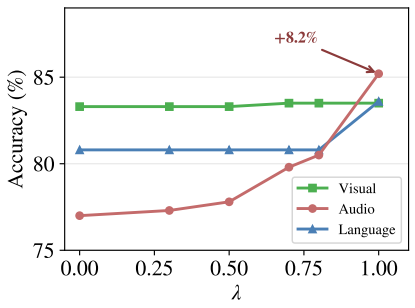
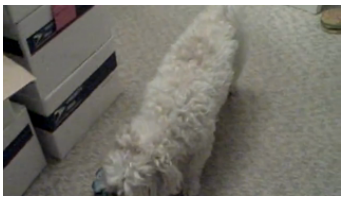


Figure 3: Per-domain accuracy on CMM with QWEN2.5-OMNI as λ increases.

V→A | AVHBench

Q: Is the vehicle making sound in the audio?

GT: No ✗Baseline: Yes ✓MGCD: No



V→A | CMM

Q: Did you hear person making sound in the audio?

GT: No ✗Baseline: Yes ✓MGCD: No

V→A | CMM

Q: Did you hear horses neighing in the audio?

GT: No ✗Baseline: Yes ✓MGCD: No

A→V | AVHBench

Q: Is the fire truck visible in the video?

GT: No ✗Baseline: Yes ✓MGCD: No

V→A | AVHBench

Q: Is the clock making sound in the audio?

GT: No ✗Baseline: Yes ✓MGCD: No

V→A | CMM

Q: Did you hear alarm clock ticking in the audio?

GT: No ✗Baseline: Yes ✓MGCD: No

Figure 4: Representative case studies. Top row: QWEN2.5-OMNI; bottom row: VIDEOLLAMA2-AV. Each case shows a sample where standard multi-branch contrastive decoding produces a hallucinated answer while MGCD corrects it by grounding on the target modality. Additional cases appear as Figure 6 in Appendix H.

Visual at 6.7% over the strongest baseline, the domain most affected by cross-modal interference, while Audio and Language are preserved at 0.9% and 0.3% respectively. The consistency across two architecturally different fusion paradigms, together with the wide margins over VCD_{Ext} and AVCD, supports that the improvement stems from the grounding mechanism rather than stronger fusion alone. The ablation in Section 5 further confirms that both components of MGCD contribute independently, with margin-level fusion alone already outperforming MAD's logit-level fusion.

Figure 3 shows the per-domain effect: as λ increases to 1.0, the Audio domain improves by 8.2% while Visual and Language remain stable. This pattern aligns with the CMI analysis: MGCD selectively corrects the domain where cross-modal interference is strongest. Note: λ=0 reflects margin-level fusion without grounding, which differs from the logit-level baseline in Table 3. McNemar's test confirms statistical significance over the strongest baseline on AVHBench for both models ($p < 0.01$).

4.3 Ablation Studies

Grounding ablation. Table 4 isolates grounding from entropy-adaptive γ on QWEN2.5-OMNI. Grounding is the dominant contributor at 1.0%; entropy-adaptive γ adds a complementary 0.1%, with a combined gain of 1.1%. The decomposition is architecture-dependent: for VIDEOLLAMA2-AV on CMM (Table 5), the margin-space contrastive baseline (λ=0, Weighted) already exceeds MAD by only 0.6%, while grounding contributes a larger 1.2%, confirming that grounding is the primary driver for modular architectures.

Table 4: Grounding ablation on AVHBench with QWEN2.5-OMNI (%).

	w/o Grounding	w/ Grounding
w/o Entropy-γ	82.3	$83.3_{\blacktriangle 1.0}$
w/ Entropy-γ	$82.4_{\blacktriangle 0.1}$	**83.4**$_{\blacktriangle 1.0}$

Fusion strategy ablation. Table 5 tests three fusion approaches on CMM with VIDEOLLAMA2-AV. MGCD yields positive gains under all three strategies, ranging from 0.3% to 1.2%, with weighted fusion benefiting most since it amplifies modality overshoot, providing more room for

Table 5: Fusion strategy ablation on CMM with VIDEOLLAMA2-AV (%).

Fusion	w/o grounding	+ MGCD
Uniform	82.3	$82.6_{\blacktriangle 0.3}$
Argmax	82.9	$83.2_{\blacktriangle 0.3}$
Weighted	82.0	$83.2_{\blacktriangle 1.2}$

Table 6: Branch weight ablation on CMM with VIDEOLLAMA2-AV (%). Disabling individual branch weights reduces both baseline accuracy and MGCD gain, but MGCD always provides positive improvement.

w_a	w_v	w_{av}	**w/o grounding**	**+ MGCD**	Δ
✗	✓	✓	81.1	82.1	+1.0
✓	✗	✓	81.9	82.9	+1.0
✓	✓	✗	82.1	82.8	+0.7
✓	✓	✓	82.0	**83.2**	**+1.2**

grounding correction. After grounding, the accuracy gap across the three strategies narrows from 0.9% to 0.6%, with Weighted reaching the same 83.2% as Argmax despite starting 0.9% lower, indicating that grounding is robust to the choice of fusion mechanism.

Branch weight ablation. Table 6 ablates branch weights on CMM with VIDEOLLAMA2-AV. Grounding improves under every configuration, with gains from 0.7% to 1.2%; the largest gain occurs when all branches are active, and positive improvement persists even when one branch is disabled. Disabling the audio branch causes the largest baseline drop to 81.1%, yet grounding still recovers 1.0% because M^* is computed independently from the audio-only branch and remains available as the correction signal. Disabling the full-input branch yields the smallest grounding gain at 0.7%, consistent with this branch contributing the most modality overshoot that grounding can correct.

4.4 Generalization to Other AVQA Benchmarks

Table 7 extends the evaluation to three general AVQA benchmarks. MGCD improves over both Vanilla and MAD on all three, with the largest gain of 4.8% over Vanilla on MUSIC-AVQA. On OmniBench, MAD degrades by 0.6% relative to Vanilla while MGCD improves by 1.5%; on WorldSense, MGCD achieves 4.6% over Vanilla and 2.3% over MAD. These gains on diverse task formats corroborate the generality of the margin-based correction.

Table 7: Generalization to other AVQA benchmarks on VIDEOLLAMA2-AV (%).

Method	**Omni.**	**World.**	**AVQA**
Vanilla	36.3	23.3	78.1
+ MAD	$35.7_{\blacktriangledown 0.6}$	$25.6_{\blacktriangle 2.3}$	$79.1_{\blacktriangle 1.0}$
+ MGCD	$\mathbf{37.8}_{\blacktriangle 1.5}$	$\mathbf{27.9}_{\blacktriangle 4.6}$	$\mathbf{82.9}_{\blacktriangle 4.8}$

5 Analysis

Oracle efficiency. We define an oracle that perfectly flips all fixable errors (samples where $M^* < 0$ and the prediction is incorrect). Table 8 shows MGCD achieves **70%** oracle efficiency on QWEN2.5-OMNI, recovering 2.8% of the 4.0% ceiling, and 50% on VIDEOLLAMA2-AV. Table 4 validates that both components contribute independently: removing grounding reduces accuracy by 1.0%, and further removing entropy-adaptive weighting drops an additional 0.1%. The correction precision of 71.8–77.8% (Table 12, Appendix I) shows that the majority of margin flips are beneficial, leaving room for further gains with refined sample selection.

Table 8: Oracle analysis on AVHBench (%). MGCD recovers 70% of the oracle ceiling on QWEN2.5-OMNI and 50% on VIDEOLLAMA2.

Configuration	**QWEN2.5**	**VIDEOLLAMA2**
Vanilla	75.5	76.8
Best baseline	80.6	79.6
MGCD	$\mathbf{83.4}_{\blacktriangle 2.8}$	$\mathbf{82.2}_{\blacktriangle 2.6}$
Oracle	$84.6_{\blacktriangle 4.0}$	$84.8_{\blacktriangle 5.2}$

6 Conclusion

We introduced MGCD, a training-free framework that anchors multi-branch contrastive decoding to the target modality's independent judgment through continuous soft correction guided by CMI

analysis. With a single hyperparameter, MGCD recovers 50–70% of the oracle ceiling on fixable errors across both end-to-end and modular architectures. The margin-based formulation is in principle applicable to other modality pairs and multi-choice settings, and we see integration with training-time calibration as a promising direction to close the remaining gap to the oracle.

References

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Conference on Empirical Methods in Natural Language Processing*, 2023a.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

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*, pages 18959–18969, 2025.

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 *13th International Conference on Learning Representations, ICLR 2025*, pages 49529–49556. International Conference on Learning Representations, ICLR, 2025.

Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. *arXiv preprint arXiv:2410.12787*, 2024a.

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

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

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B 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)*, pages 12286–12312, 2023b.

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*, pages 13872–13882, 2024b.

Chaeyoung Jung, Youngjoon Jang, and Joon Son Chung. Avcd: Mitigating hallucinations in audio-visual large language models through contrastive decoding. *arXiv preprint arXiv:2505.20862*, 2025.

Sangyun Chung, Se Yeon Kim, Youngchae Chee, and Yong Man Ro. Mad: Modality-adaptive decoding for mitigating cross-modal hallucinations in multimodal large language models. *arXiv preprint arXiv:2601.21181*, 2026.

Baolong Bi, Shenghua Liu, Lingrui Mei, Yiwei Wang, Junfeng Fang, Pengliang Ji, and Xueqi Cheng. Decoding by contrasting knowledge: Enhancing large language model confidence on edited facts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17198–17208, 2025.

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*, pages 15840–15853, 2024.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023c.

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*, pages 13418–13427, 2024.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer, 2024.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: object hallucination reduction via adaptive focal-contrast decoding. In *Proceedings of the 41st International Conference on Machine Learning*, pages 7824–7846, 2024.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14303–14312. IEEE Computer Society, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pages 24043–24055. PMLR, 2022.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8238–8247, 2022.

Yake Wei and Di Hu. Mmpareto: boosting multimodal learning with innocent unimodal assistance. In *Proceedings of the 41st International Conference on Machine Learning*, pages 52559–52572, 2024.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pages 41753–41769. PMLR, 2023b.

Bing Cao, Yinan Xia, Yi Ding, Changqing Zhang, and Qinghua Hu. Predictive dynamic fusion. *arXiv preprint arXiv:2406.04802*, 2024.

Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26866–26875, 2024. URL `https://api.semanticscholar.org/CorpusID:272722535`.

Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability, volume 1: Contributions to the theory of statistics*, volume 3, pages 197–207. University of California Press, 1956.

William James, Charles Stein, et al. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379. University of California Press, 1961.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

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

Yizhi Li, Yinghao Ma, Ge Zhang, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, et al. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*, 2024.

Youssef Benchekroun, Megi Dervishi, Mark Ibrahim, Jean-Baptiste Gaya, Xavier Martinet, Grégoire Mialon, Thomas Scialom, Emmanuel Dupoux, Dieuwke Hupkes, and Pascal Vincent. Worldsense: A synthetic benchmark for grounded reasoning in large language models. *arXiv preprint arXiv:2311.15930*, 2023.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19108–19118, 2022.

Table 9: Notation used throughout the paper.

Symbol	Definition	Description
Input and model		
$x = (x_v, x_a, x_t)$	Multi-modal input	Visual, audio, and text components
l_{av}, l_v, l_a, l_t	Branch logit vectors	Full, visual-only, audio-only, text-only branches
y_1, y_2	Candidate answers	y_1: higher-scoring option under l_{av}; y_2: second option
Margins		
M_i	$l_i(y_1) - l_i(y_2)$	Branch margin for branch i, see Eq. 1
M^*	Target modality margin	M_a for V→A tasks, M_v for A→V tasks, see Eq. 1
M_f	$\sum_i w_i(x) \cdot M_i$	Fused margin, weighted combination of branch margins
M_g	Modality-grounded margin	Core output of MGCD, see Eq. 5
Fusion weights		
$p_i(x)$	Modality relevance	Model-estimated relevance of modality i
$w_i(x)$	$\gamma \cdot p_i(x)$	Fusion weight for branch i; Algorithm 1 uses γ_i
γ	Contrastive strength	Base scaling factor for fusion weights
w_t	-1	Contrastive weight for text branch
q_i	Output distribution of branch i	Over K candidate tokens; used for entropy-adaptive γ
γ_i	$\gamma \cdot (1 - H(q_i)/\log K)$	Entropy-adaptive contrastive strength for branch i
K	Candidate answer token count	e.g., 2 for Yes/No tasks
d	Task direction	$d \in \{\text{V}\to\text{A}, \text{A}\to\text{V}\}$
ϕ	Directional activation flag	If true, grounding is restricted to V→A direction
Cross-modal interference		
C_v, C_a, C_I	Modality contributions	Visual, audio, and interaction terms, see Eq. 2
C_i^M	$C_i(y_1) - C_i(y_2)$	Margin-level modality contribution
CMI	$\frac{\lvert C^M_{\text{non-target}}\rvert}{\lvert C^M_{\text{non-target}}\rvert + \lvert C^M_{\text{target}}\rvert}$	Cross-Modal Interference ratio $\in [0, 1]$, see Eq. 3
Signal model and MGCD		
y^*	$\in \{+1, -1\}$	True label
δ_f, δ^*	Signal strengths	Of fused and target margins respectively
σ_f, σ^*	Noise std. devs.	Of fused and target margins; $\sigma_f > \sigma^*$
r	δ^*/δ_f	Signal ratio, target-to-fused
s	σ^*/σ_f	Noise ratio; $s < 1$ guaranteed
ρ	$\text{Corr}(\epsilon_f, \epsilon^*)$	Noise correlation; $\rho \geq 0$ in practice, see Corollary 3
λ	Anchoring strength	Controls grounding intensity ($\lambda > 0$); starting point $\gamma/3$
$\lambda_{\max}$	See Proposition 1	Maximum anchoring strength yielding positive SNR gain
Additional notation		
f	MLLM	The multimodal large language model
$\varnothing$	Null modality input	Masked modality placeholder
$H(\cdot)$	Shannon entropy	Used in entropy-adaptive γ_i
$\hat{y}$	Prediction	y_1 if $M_g > 0$, else y_2
$\mathcal{O}$	$\{M_f > M^*\}$	Overshoot subset where correction is applied
$\mathcal{O}^c$	$\{M_f \leq M^*\}$	Complement; $M_g = M_f$ (no correction)

A Notation

Table 9 summarizes all symbols used throughout the paper, organized into six semantic groups: input and model, margins, fusion weights, cross-modal interference, the signal model underlying MGCD, and additional notation. We follow the convention that positive margins indicate agreement with the higher-scoring candidate y_1 under the full branch l_{av}, and negative margins indicate disagreement.

Precise characterization and margin-level expansion of modality contributions. On the decomposition in Eq. 2. The additive decomposition in Eq. 2 is more precisely described as the *inclusion-exclusion (Möbius) decomposition* of the cooperative game between modalities, not the Shapley value itself. The Shapley value for the visual modality in a two-player game would be $\phi_v = l_t + \frac{1}{2}C_v + \frac{1}{2}C_I$ (allocating half the interaction term to each player), which is a different quantity than C_v as used in Eq. 2. The term "Shapley-based attribution" in the main text refers to

the broader framework of cooperative game attribution [Lundberg and Lee, 2017]; the exact form used is the Möbius/Harsanyi dividend decomposition, which treats the interaction C_I as a separate term rather than distributing it among players. CMI uses only the single-modality terms C_v^M and C_a^M (excluding C_I^M) to isolate per-modality contributions without allocating the interaction.

Closed-form expansion. The margin-level counterpart $C_i^M = C_i(y_1) - C_i(y_2)$ has a compact closed form. Substituting $C_v(y) = l_v(y) - l_t(y)$:

$$\begin{aligned} C_v^M &= C_v(y_1) - C_v(y_2) \\ &= \big[l_v(y_1) - l_t(y_1)\big] - \big[l_v(y_2) - l_t(y_2)\big] \\ &= \big[l_v(y_1) - l_v(y_2)\big] - \big[l_t(y_1) - l_t(y_2)\big] = M_v - M_t. \end{aligned} \tag{10}$$

By the same argument, $C_a^M = M_a - M_t$, and $C_I^M = M_{av} - M_v - M_a + M_t$. These identities show that the CMI ratio in Eq. 3 depends only on the four branch margins already computed by the MGCD pipeline.

Causal justification for $\sigma_f > \sigma^*$. The assumption $\sigma_f > \sigma^*$ in Section 3.2 is motivated by the CMI analysis as follows. High-CMI samples have substantially higher error rates (Figure 2(a)), meaning the non-target modality's contribution systematically pushes the fused prediction toward the wrong answer. Under the signal model of Eq. 4, this manifests as an increase in the variance of the fused margin M_f relative to the target-only margin M^*: the non-target contribution adds variance (noise) without adding proportional signal, since it is oriented toward a biased prediction rather than the true label. Formally, if the non-target logit difference $C^M_{\text{non-target}}$ is independent of the true label y^* on high-CMI samples (pure noise), then $\text{Var}(M_f) > \text{Var}(M^*)$, confirming $\sigma_f > \sigma^*$. The empirical error-rate pattern (a 19–39.5% gap between low- and high-CMI bins) is consistent with this interpretation.

Approximation in the one-sided gating argument (Remark 2). The argument in Remark 2 that the one-sided gate preserves the SNR improvement from Proposition 1 is an *approximation*. Proposition 1 analyzes the unconditional SNR over all samples; Remark 2 implicitly transfers this analysis to the conditional subset $\mathcal{O} = \{M_f > M^*\}$. Conditioning on $\mathcal{O}$ may shift the effective signal strengths δ_f, δ^* and noise standard deviations σ_f, σ^* relative to the unconditional case, so the proposition's guarantees do not strictly carry over. The two-case analysis in the remark (correct predictions preserved; incorrect predictions corrected) is heuristic, relying on the empirical observation that $M^* < 0$ with high probability when $y^* = -1$ (Table 10). The empirical correction precision of 71.8–77.8% (Table 12) and the consistent accuracy gains in Section 4.2 provide the primary validation that the gate behaves as described in practice.

B Modality Relevance Prompting

The fusion weights $w_i(x) = \gamma \cdot p_i(x)$ depend on the model's self-assessed modality relevance probabilities $p_i(x)$. We obtain these probabilities by prepending a short relevance query before the main question, a strategy also used in recent multi-branch contrastive methods [Jung et al., 2025, Chung et al., 2026]. Specifically, the model is prompted with: "To answer the following question, which modalities are needed? Options: audio, video, both." We extract the first-token logits over the three keyword tokens and apply softmax to obtain $p_a(x)$, $p_v(x)$, and $p_{av}(x)$. The text branch always receives a fixed contrastive weight $w_t = -1$, independent of the relevance query. This prompting step adds a single short forward pass per sample. On both QWEN2.5-OMNI and VIDEOLLAMA2-AV, the model assigns higher relevance to the modality that is genuinely needed for the question in the majority of cases, which enables the dynamic weighting to adapt per sample.

C Target Margin Bucketing

Table 10 shows the same declining trend on VIDEOLLAMA2-AV as Figure 2(b) on QWEN2.5-OMNI. When M^* is strongly negative, i.e., the target modality disagrees with the fused prediction, the error rate reaches 72.0% in the $[-5, -2)$ range. As M^* increases, the error rate drops to 50.4% in the $[-2, 0)$ range, then falls sharply to 16.6% for $[0, 2)$ and 5.7% for $[2, 5)$. The steepest transition occurs around $M^* = 0$, where the error rate drops by more than 30%; the sign of M^* is thus a strong

Table 10: Error rate by target modality margin M^* (VideoLLaMA2, V→A + pred=Yes, N=1170). Negative M^* indicates high error risk.

M^* **range**	N	**Error rate**
$[-5, -2)$	50	72.0%
$[-2, 0)$	244	50.4%
$[0, 2)$	349	16.6%
$[2, 5)$	475	5.7%
$[5, 10)$	52	3.8%

indicator of prediction correctness. This pattern mirrors the QWEN2.5-OMNI results and confirms that M^* serves as a reliable per-sample error indicator across both architectures, supporting the use of M^* as the anchoring signal in MGCD. The consistency of this trend across an end-to-end model (QWEN2.5-OMNI) and a modular model (VIDEOLLAMA2-AV) suggests that the phenomenon is not an artifact of a particular architecture but a general property of how multi-branch contrastive decoding distributes confidence across modality-specific branches.

D λ Robustness

MGCD introduces a single additional hyperparameter λ, the anchoring strength, with $\lambda = \gamma/3$ as the starting point. Figure 5 shows the full λ sweep on AVHBench for both models. QWEN2.5-OMNI exhibits a plateau over $\lambda \in [0.7, 1.0]$ where accuracy varies by less than 0.7%, with a sharp cliff at $\lambda = 1.2$ when the correction overshoots the target margin. VIDEOLLAMA2-AV shows a similar pattern at a smaller scale: a plateau over $\lambda \in [0.1, 0.3]$ followed by gradual decline and a cliff at $\lambda = 0.9$. In both cases, $\gamma/3$ falls within the effective plateau, and all λ values in the plateau improve over the ungrounded baseline.

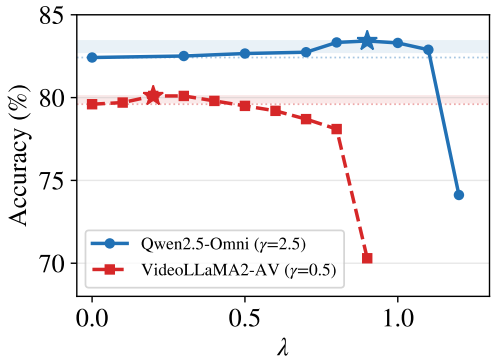


Figure 5: λ sweep on AVHBench without directional activation. Both models exhibit a performance plateau around $\gamma/3$, followed by a cliff when λ exceeds the convex combination regime. Stars mark the optimal λ.

E Proof of Proposition 1 and Correlated Noise Extension

Proof of Proposition 1. Under the signal model (Eq. 4) with independent noise, $M_g = (1-\lambda)M_f + \lambda M^*$ has signal $(1-\lambda)\delta_f + \lambda\delta^*$ and variance $(1-\lambda)^2\sigma_f^2 + \lambda^2\sigma^{*2}$. The condition $\text{SNR}(M_g) > \text{SNR}(M_f) = \delta_f/\sigma_f$ is equivalent to (squaring both sides and cross-multiplying by the positive denominators):

$$\sigma_f^2\left[(1-\lambda)\delta_f + \lambda\delta^*\right]^2 > \delta_f^2\left[(1-\lambda)^2\sigma_f^2 + \lambda^2\sigma^{*2}\right]. \tag{11}$$

Expanding and canceling the common $(1-\lambda)^2\delta_f^2\sigma_f^2$ term on both sides, then dividing by $\lambda\delta_f^2\sigma_f^2 > 0$, yields:

$$2(1-\lambda)\, r + \lambda(r^2 - s^2) > 0, \tag{12}$$

where $r = \delta^*/\delta_f$ and $s = \sigma^*/\sigma_f < 1$.

Case (i): $r \geq s$. For any $\lambda \in (0, 1)$: $2(1-\lambda)r > 0$ since $r > 0$ and $1-\lambda > 0$; $\lambda(r^2 - s^2) \geq 0$ since $r \geq s$. Their sum is strictly positive, so $\lambda_{\max} = 1$.

Case (ii): $r < s$. The term $r^2 - s^2 < 0$, so the left-hand side of Eq. 12 is a decreasing affine function of λ. At $\lambda = 0$, it evaluates to $2r > 0$ (satisfied since $r > 0$). Setting Eq. 12 to zero gives:

$$\lambda_{\max} = \frac{2r}{2r + s^2 - r^2}.$$

Since $r > 0$, the numerator is positive; since $s > r$, the denominator satisfies $2r + s^2 - r^2 > 2r$, so $\lambda_{\max} \in (0, 1)$. □

We now relax the independence assumption. Since M_f includes M^* as a weighted component, the noise terms ϵ_f and ϵ^* are positively correlated in practice. Let $\rho = \text{Corr}(\epsilon_f, \epsilon^*) \in [0, 1)$.

Corollary 3 (Effective interval under correlated noise)**.** *Let $\rho = Corr(\epsilon_f, \epsilon^*) \in [0, 1)$. If $r > \rho\, s$ (the target modality's signal advantage exceeds the correlation-induced redundancy), there exists $\lambda_{\max}(\rho) > 0$ such that SNR$(M_g) >$ SNR(M_f) for all $\lambda \in (0, \lambda_{\max}(\rho))$:*

$$\lambda_{\max}(\rho) = \begin{cases} 1 & \text{if } r \geq s, \\ \dfrac{2(r - \rho\, s)}{2(r - \rho\, s) + s^2 - r^2} & \text{if } \rho\, s < r < s. \end{cases} \tag{13}$$

When $\rho = 0$, this reduces to Proposition 1. Positive correlation strictly shrinks the effective interval: $\lambda_{\max}(\rho) < \lambda_{\max}(0)$ whenever $\rho > 0$ and $r < s$.

Proof. With $\text{Cov}(\epsilon_f, \epsilon^*) = \rho\, \sigma_f \sigma^*$, the variance of M_g becomes $(1-\lambda)^2\sigma_f^2 + \lambda^2\sigma^{*2} + 2\lambda(1-\lambda)\rho\, \sigma_f \sigma^*$. Repeating the SNR comparison from Proposition 1 (squaring, cross-multiplying, expanding, and dividing by $\lambda\delta_f^2\sigma_f^2 > 0$) yields the generalized condition:

$$2(1-\lambda)(r - \rho\, s) + \lambda(r^2 - s^2) > 0. \tag{14}$$

This has identical structure to Eq. 12 with r replaced by $r - \rho\, s$.

Case $r \geq s$: Since $\rho < 1$, we have $\rho\, s < s \leq r$, so $r - \rho\, s > 0$. Both terms in Eq. 14 are non-negative for $\lambda \in (0, 1)$, giving $\lambda_{\max}(\rho) = 1$.

Case $\rho\, s < r < s$: The intercept $2(r - \rho\, s) > 0$ and the slope $r^2 - s^2 - 2(r - \rho\, s) < 0$ (since $r^2 - s^2 < 0$ and $r - \rho\, s > 0$). Setting Eq. 14 to zero gives $\lambda_{\max}(\rho) = 2(r - \rho\, s)/[2(r - \rho\, s) + s^2 - r^2]$. Since increasing ρ decreases the numerator while leaving the addend $s^2 - r^2 > 0$ unchanged, $\lambda_{\max}(\rho)$ is strictly decreasing in ρ. □

Interpretation. The condition $r > \rho\, s$ states that the target branch must provide information beyond what is already captured in the fused margin through correlation. In the limiting case $\rho \to 0$ (independent noise), any positive signal ratio $r > 0$ suffices. As ρ increases, a stronger target signal is required to overcome the redundancy, and the safe range for λ narrows. In practice, the moderate values of ρ induced by the weighted fusion leave substantial room for improvement, as confirmed by the consistent gains in Section 4.2.

F Hyperparameter Settings

Table 11: Hyperparameter settings per model and dataset.

Model	**Dataset**	γ	λ
Qwen2.5	AVHBench	2.5	0.9
Qwen2.5	CMM	2.5	1.0
VideoLLaMA2	AVHBench	0.5	0.2
VideoLLaMA2	CMM	0.5	0.3

Table 11 lists all hyperparameter values used in the main evaluation. We set the contrastive strength $\gamma = 2.5$ for QWEN2.5-OMNI and $\gamma = 0.5$ for VIDEOLLAMA2-AV. The five-fold difference reflects the architectures' different logit scales: QWEN2.5-OMNI produces larger raw logit magnitudes, requiring a proportionally larger contrastive weight to achieve a comparable contrastive effect. The anchoring strength λ is the only new hyperparameter introduced by MGCD; it is selected by grid search over $[0.0, 2.0]$ in steps of 0.1 on a held-out validation split from each dataset. As discussed in Section D, $\gamma/3$ provides a reliable starting point: for QWEN2.5-OMNI, $\gamma/3 \approx 0.83$ falls within the plateau $[0.7, 1.0]$; for VIDEOLLAMA2-AV, $\gamma/3 \approx 0.17$ falls within the plateau $[0.1, 0.3]$. The ratio between the optimal λ and γ ranges from 0.36 to 0.60 across the four model-dataset configurations. Three of the four settings fall in $[0.36, 0.40]$, close to $\gamma/3 \approx 0.33$; the outlier is VIDEOLLAMA2-AV on CMM ($\lambda/\gamma = 0.60$), where the modular architecture's stronger per-modality separation may

warrant a larger anchoring strength. Despite this variation, all optimal λ values lie within the effective plateau identified in Figure 5, and the $\gamma/3$ starting point remains a practical default.

For all general AVQA benchmarks in Table 7, we use the same γ as the model's primary setting and $\lambda = 0.1$. The conservative λ for general benchmarks reflects the fact that these tasks are not specifically designed to test cross-modal hallucination, so a lighter correction avoids interfering with the model's general reasoning ability. No task-specific tuning is performed on these benchmarks; the same $\lambda = 0.1$ is applied uniformly across OmniBench, WorldSense, and MUSIC-AVQA.

G Dataset Details

We evaluate on two hallucination-specific benchmarks, AVHBench and CMM, and three general audio-visual QA benchmarks, OmniBench, WorldSense, and MUSIC-AVQA. Below we describe the sub-category structure and evaluation metrics for each.

AVHBench. AVHBench [Sung-Bin et al., 2025] is a cross-modal hallucination benchmark consisting of 3,419 Yes/No questions in two directional subsets: V→A with 2,287 samples, which asks whether a sound described in the question is present in the audio, and A→V with 1,132 samples, which asks whether an object implied by the audio is visible in the video. The V→A subset is roughly twice as large as A→V, reflecting the greater prevalence of visual-to-audio interference scenarios in naturalistic video data. Because of this size imbalance, the AVHBench average reported in Table 3 is sample-weighted rather than a simple mean of the two direction-level accuracies.

CMM. CMM [Leng et al., 2024a] evaluates modality over-reliance through 2,400 Yes/No questions spanning five subcategories. We evaluate on the three cross-modal domains: Visual, which tests over-reliance on audio while ignoring visual cues; Audio, which tests over-reliance on visual input while ignoring audio; and Language, which tests over-reliance on language priors while ignoring visual evidence. The remaining two subcategories test non-cross-modal biases and are excluded from our evaluation, leaving 1,200 samples. Each domain contains 200 positive and 200 negative samples. The domain-level accuracy is computed as the average of presence accuracy and hallucination rejection, and the CMM average reported in Table 3 is the simple mean across the three domains.

General AVQA benchmarks. OmniBench [Li et al., 2024] is a multiple-choice benchmark requiring joint reasoning over visual, audio, and textual inputs. WorldSense [Benchekroun et al., 2023] tests grounded spatial reasoning through multiple-choice questions with synthetic scene descriptions. MUSIC-AVQA [Li et al., 2022] contains over 45K open-ended questions about dynamic audio-visual scenes, requiring spatio-temporal grounding of sounds and visual objects. These three benchmarks are used to validate generalization beyond hallucination-specific tasks, as reported in Table 7.

H Additional Case Studies

Figure 4 presents six representative cases in the main text. Here we show six additional examples covering the remaining hallucination types and dataset combinations.

I Correction Precision

Table 12: Correction precision on full AVHBench (%). TF = true fixes, NE = new errors.

Model	TF	NE	Precision
QWEN2.5-OMNI	56	22	71.8%
VIDEOLLAMA2-AV	119	34	77.8%

As shown in Table 12, all corrections occur exclusively on samples where $M^* < 0$, with precision 71.8–77.8% across both models, meaning the majority of prediction changes are beneficial. The average CMI of corrected samples is 0.545 on QWEN2.5-OMNI, compared to 0.48 in the activation zone overall, consistent with MGCD preferentially fixing high-interference samples. A detailed AUC-ROC analysis comparing the predictive power of different signals is in Appendix J.

V→A | AVHBench

Q: Is the race track making sound in the audio?

GT: No ✗Baseline: Yes ✓MGCD: No

A→V | AVHBench

Q: Is the vehicle visible in the video?

GT: Yes ✗Baseline: No ✓MGCD: Yes

V→A | CMM

Q: Did you hear motorcycle revving in the audio?

GT: No ✗Baseline: Yes ✓MGCD: No

A→V | AVHBench

Q: Are the birds visible in the video?

GT: No ✗Baseline: Yes ✓MGCD: No

V→A | CMM

Q: Did you hear motorcycle revving in the audio?

GT: No ✗Baseline: Yes ✓MGCD: No

V→A | CMM

Q: Did you hear fan blowing in the audio?

GT: No ✗Baseline: Yes ✓MGCD: No

Figure 6: Additional case studies. Top row: QWEN2.5-OMNI; bottom row: VIDEOLLAMA2-AV. The same pattern from Figure 4 holds: standard contrastive decoding hallucinates cross-modal content while MGCD anchors to the target modality.

J Signal Predictive Power

Table 13: AUC-ROC comparison of candidate signals in the V→A + pred=Yes activation zone. Higher AUC indicates better error prediction.

Signal	VideoLLaMA2	Qwen2.5
M^* (target margin)	**0.821**	0.827
M_f (fused margin)	0.780	**0.856**
$\lvert M_f - M^* \rvert$	0.703	0.853
−CMI (low=safe)	0.594	0.648

Table 13 compares the predictive power of various signals for classifying correct and incorrect predictions in the activation zone, defined as V→A samples where the model predicts Yes. We use AUC-ROC as the evaluation metric, where each signal is treated as a continuous score and the binary label is whether the prediction is correct. The activation zone is chosen because it concentrates the majority of fixable errors: in V→A tasks where the model predicts Yes, the fused margin is positive, and a negative M^* signals that the target modality disagrees with this prediction.

The target margin M^* achieves AUC ≥ 0.82 on both models, far exceeding CMI at 0.59–0.65, which shows that M^* carries substantially more per-sample information than the population-level CMI ratio. The gap $|M_f - M^*|$ also achieves high AUC of 0.853 on QWEN2.5-OMNI, showing that the overshoot magnitude itself is predictive of errors. On QWEN2.5-OMNI, M_f attains a higher AUC of 0.856 compared to 0.827 for M^*; this is unsurprising because M_f already determines the prediction and therefore trivially predicts its own correctness. MGCD does not replace M_f with M^* but uses the directional disagreement $M_f - M^*$ to correct overshoot, so M^*'s value as an anchor signal is complementary to M_f's predictive power. On VIDEOLLAMA2-AV the pattern reverses: M^* achieves 0.821 while M_f reaches only 0.780, suggesting that the target modality provides a stronger error signal on the modular architecture. This cross-architecture difference is consistent with the observation that modular models maintain more independent per-modality representations, making the target branch a more reliable standalone signal.

K Computational Resources and Evaluation Protocol

Hardware and runtime. All experiments are conducted on a server equipped with 8×NVIDIA H20 GPUs (96 GB HBM each). Each evaluation run requires a single GPU. For AVHBench (3,419 samples), a single forward pass through all four branches takes approximately 2.5 hours on QWEN2.5-OMNI and 1.5 hours on VIDEOLLAMA2-AV. CMM (1,200 samples) completes in roughly 1 hour per model. The λ sweep (21 values from 0.0 to 2.0) reuses cached branch logits and adds negligible overhead, since only the margin computation and grounding step are repeated. The total compute for all experiments reported in this paper, including λ sweeps, ablations, and general AVQA benchmarks, is approximately 120 GPU-hours.

Inference overhead. MGCD requires four forward passes per sample (full, visual-only, audio-only, text-only), compared to one for vanilla decoding. The additional overhead is purely in the forward pass stage; the margin computation, CMI calculation, and grounding step together take less than 1 ms per sample on CPU and are negligible relative to the model inference time. The modality relevance query that produces $p_i(x)$ is a single additional forward pass with a short prompt.

Evaluation protocol. For Yes/No tasks on AVHBench and CMM, we extract the first generated token and classify based on whether it starts with "Yes" or "No". When the model generates a token that does not match either keyword, we fall back to the higher-logit candidate between the Yes and No tokens; this fallback is triggered on fewer than 0.5% of samples across all configurations. For multiple-choice tasks on OmniBench and WorldSense, we compare logits of the candidate option tokens A, B, C, and D and select the highest. For MUSIC-AVQA, we follow the standard GPT-based evaluation protocol, where a language model judges whether the generated answer matches the ground truth.

L Broader Impacts

MGCD reduces cross-modal hallucinations in multimodal large language models without additional training, lowering the barrier to deploying more reliable AI systems in safety-sensitive domains such as medical video analysis and accessibility tools. Reduced hallucination rates can increase user trust; however, this trust should be calibrated against the method's remaining error rate and the domains covered by the evaluation benchmarks. The training-free nature allows MGCD to be applied to models where retraining is infeasible, broadening practical reach without introducing new data privacy concerns.

M Limitations

We evaluate on two models spanning end-to-end and modular paradigms; broader model coverage would further validate the $\lambda = \gamma/3$ heuristic. MGCD requires four forward passes per sample, comparable to other multi-branch contrastive methods, but approximately 5× over standard decoding. We validate on audio-visual cross-modal hallucinations; other modality combinations remain unexplored. The top-two reduction in the multiple-choice extension means MGCD cannot recover when the correct answer ranks third or lower under the fused logits. The current λ is set via grid search; an adaptive scheme that uses per-sample overshoot magnitude or CMI to modulate λ could remove the need for any hyperparameter tuning. MGCD cannot correct errors where the target modality margin itself is wrong ($M^* \geq 0$ but the prediction is incorrect), and correction-induced errors account for the gap between MGCD and the oracle ceiling.