MITIGATING CROSS-MODAL HALLUCINATIONS BY CORRECTING MODALITY MISROUTING IN OMNI-MODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Omnimodal large language models (OLLMs) are susceptible to cross-modal hallucinations under audio-visual interference. We find that, for a substantial fraction of incorrect joint predictions, the question-relevant unimodal branch can still recover the correct answer, suggesting that task-relevant evidence can remain recoverable from the relevant modality even when it is not reflected in the joint prediction. We refer to this recoverable failure pattern as *modality misrouting*. We propose **CaRe (Conflict-aware Residual Intervention)**, a training-free inference-time method for identifying and correcting modality misrouting. CaRe uses a text-only modality probe to identify a candidate modality-specific counterfactual branch, and flags potential misrouting when its prediction conflicts with the original joint prediction. For such cases, CaRe computes a residual direction from the discrepancy between the counterfactual and fused representations and uses it to steer the original multimodal state toward the relevant modality evidence, while retaining both modalities in the original audio-visual inference path. The intervention strength is adapted based on the confidence of the counterfactual branch. Experiments on CMM and AVHBench across two OLLMs show that CaRe consistently outperforms the evaluated inference-time hallucination-mitigation baselines.

1 INTRODUCTION

Omnimodal large language models (OLLMs) jointly process audio, visual, and textual signals, enabling cross-modal perception and reasoning (Xu et al., 2025a; Yao et al., 2025; Zhang et al., 2023; Fu et al., 2026; Luo et al., 2026). However, joint multimodal modeling does not necessarily ensure faithful grounding. OLLMs may still produce responses that are not faithfully grounded in the available audio-visual evidence, leading to audio-visual hallucinations that limit their reliability in video understanding, audio-visual question answering, and embodied applications (Leng et al., 2026; Sung-Bin et al., 2025).

We investigate a recoverable form of cross-modal hallucination in which one modality interferes with evidence from the modality actually required by the question. As illustrated in Figure 1, a visible cat can cause an OLLM to falsely report hearing a meow even when the audio contains only duck quacking; conversely, birdsong can induce a hallucinated bird in the visual scene. Importantly, these errors do not necessarily imply that the question-relevant evidence has been lost: for a substantial fraction of incorrect

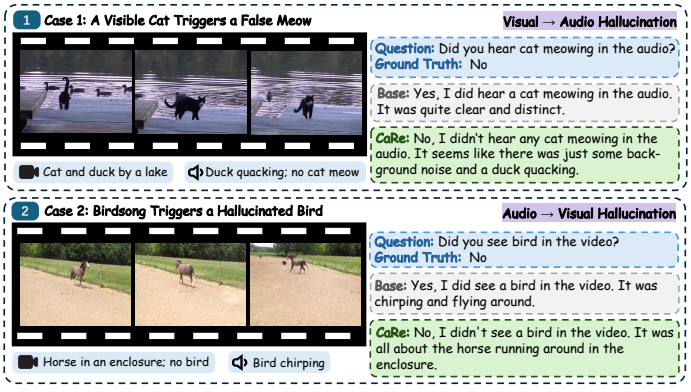

Figure 1: **Bidirectional modality misrouting under audio-visual interference.** CaRe mitigates these errors by steering the fused representation toward the selected modality-specific branch.

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.

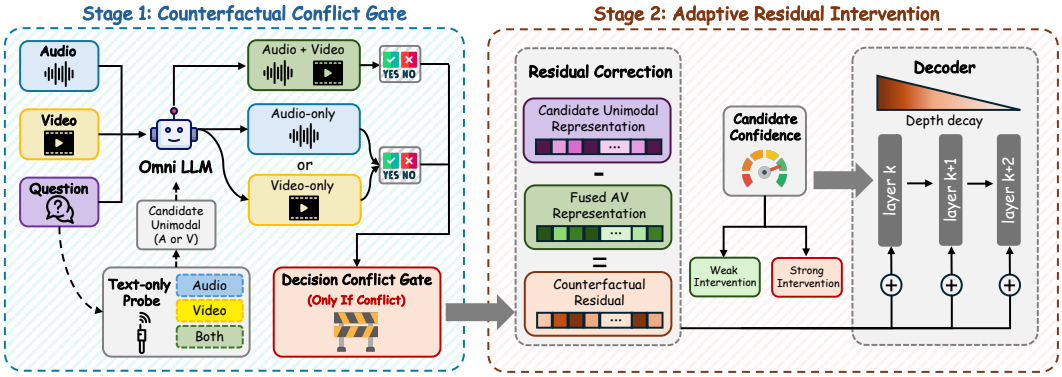

Figure 2: **Overview of CaRe.** Given audio-visual inputs and a question, CaRe uses a text-only modality probe to select a candidate modality, detects decision conflicts between the joint and counterfactual branches, and applies confidence-adaptive residual intervention at selected decoder layers with strengths decayed across depth.

joint predictions, the corresponding unimodal branch still recovers the correct answer. We refer to this discrepancy between the erroneous joint prediction and the recoverable question-relevant modality-specific prediction as *modality misrouting*. This observation suggests that some audio-visual hallucinations may be corrected by redirecting the fused representation toward evidence that remains accessible within the model.

To address this failure, we propose **CaRe (Conflict-aware Residual Intervention)**, a training-free inference-time method for selectively correcting modality misrouting. As illustrated in Figure 2, CaRe follows a two-stage pipeline for detecting potential misrouting and correcting the corresponding fused representation. First, a text-only modality probe identifies the modality most likely required by the question and selects the corresponding counterfactual modality-specific branch by withholding the competing modality. A counterfactual conflict gate then compares its prediction with the original joint prediction and activates correction only when the two exhibit a hard decision conflict. For gate-positive samples, CaRe uses counterfactual residuals to steer the fused representation toward selected modality-specific reference, with strength adapted to counterfactual confidence. Importantly, the correction is applied to the original audio-visual representation while retaining both modalities. Experiments on CMM and AVHBench across two OLLMs show that CaRe consistently outperforms the evaluated inference-time hallucination-mitigation baselines. The main contributions of this work are summarized as follows:

- We identify and characterize *modality misrouting*, an underexplored yet recoverable failure mode of OLLMs in which question-relevant modality-specific evidence remains accessible but is not effectively reflected in the joint multimodal prediction. This perspective highlights a previously underexplored failure pattern in multimodal evidence integration that can give rise to cross-modal hallucinations.
- We propose CaRe, a training-free inference-time intervention framework that can be directly applied to existing open-source OLLMs without retraining or modifying model parameters. CaRe detects potential modality misrouting through text-only modality probing and counterfactual conflict gating, and corrects it by steering the fused representation with residual signals derived from the selected modality-specific counterfactual branch.
- We evaluate CaRe on two OLLMs and two audio-visual hallucination benchmarks. Results show that CaRe consistently outperforms strong inference-time baselines while preserving the complete audio-visual input in the multimodal inference path.

2 Related Work

2.1 Omnimodal Large Language Models

Recent multimodal foundation models have extended language models toward unified processing of audio, visual, and textual signals, including OneLLM (Han et al., 2024), VITA (Fu et al., 2026),

Qwen-Omni and MiniCPM-o (Xu et al., 2025a; Yao et al., 2025; Team, 2026; Xu et al., 2025b; Cui et al., 2026), and OmniVinci (Ye et al., 2026). Despite increasingly strong joint perception, broader modality coverage does not necessarily ensure reliable evidence integration, as competing or imbalanced modalities may prevent question-relevant information from being effectively reflected in the joint prediction (Park et al., 2025; Zhang et al., 2025). Our work focuses on this inference-time failure and uses recoverable modality-specific evidence to correct erroneous multimodal predictions without modifying the underlying OLLM.

2.2 Hallucination Mitigation for Omnimodal Models

Cross-modal hallucination remains a major reliability challenge for multimodal and omnimodal models (Zechen et al., 2024; Yang et al., 2026), as highlighted by benchmarks such as CMM (Leng et al., 2026) and AVHBench (Sung-Bin et al., 2025). Existing mitigation methods mainly rely on training-time alignment (Chen et al., 2026; Chaubey et al., 2026; Yu et al., 2024; Baid et al., 2026; Pan et al., 2026; Jiang et al., 2024), representation intervention (Li et al., 2023; Chen et al., 2025; Rimsky et al., 2024; Su et al., 2025; Li et al., 2025; Wu et al., 2025; Yin et al., 2026), or training-free decoding correction (Leng et al., 2024; Jung et al., 2026; Chung et al., 2026; Favero et al., 2024; Huang et al., 2024; Ma et al., 2026), including contrastive decoding, modality reweighting, and activation steering. These approaches primarily suppress hallucinated outputs or recalibrate modality contributions, while the recoverable setting where question-relevant unimodal evidence remains available but is not effectively reflected in the joint prediction has received less attention. CaRe targets this setting by detecting counterfactual disagreement and selectively correcting the joint representation while retaining the original audio-visual inference path.

3 Motivation

Audio-visual hallucinations do not always coincide with a failure of the question-relevant unimodal view. In some cases, the joint audio-visual branch produces an incorrect prediction, while the branch retaining only the modality relevant to the question recovers the correct answer. We refer to this recoverable discrepancy between the joint and modality-specific predictions as modality misrouting. This observation suggests that some multimodal failures may still retain useful modality-specific evidence that can be exploited at inference time.

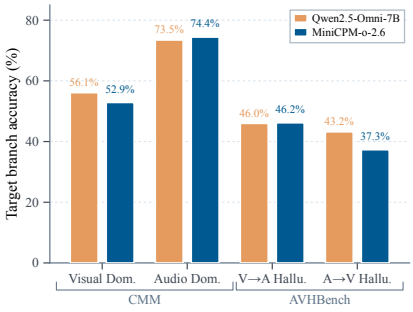

Figure 3: Target-branch recovery among incorrect joint predictions on CMM and AVHBench (%).

To quantify how often this recoverable pattern occurs, we analyze the Visual and Audio Dominance subsets of CMM together with the two cross-modal hallucination tasks of AVHBench. For each incorrect joint audio-visual prediction, we withhold the competing modality and test whether the question-relevant unimodal branch recovers the ground-truth answer. As shown in Figure 3, the recovery rate ranges from 52.9% to 74.4% on CMM and from 37.3% to 46.2% on AVHBench. The consistent recovery across both backbones, benchmarks, and interference directions shows that a nontrivial fraction of joint prediction failures remain recoverable from modality-specific evidence, motivating targeted correction of the joint representation.

4 Preliminaries

Input configurations. Given an audio input a, a visual input v, and a textual question q, we consider three input configurations:

$$x^{AV} = (a, v, q), \qquad x^{A} = (a, \varnothing_V, q), \qquad x^{V} = (\varnothing_A, v, q),$$

where $\varnothing_A$ and $\varnothing_V$ denote withheld audio and visual inputs, respectively. Running the same frozen OLLM under these configurations induces a joint audio-visual branch $\mathcal{B}_{AV}$, an audio-only branch $\mathcal{B}_A$, and a video-only branch $\mathcal{B}_V$. We regard the two unimodal branches as counterfactual views of the original audio-visual input, obtained by withholding the competing modality.

Task-level decisions. Let $\mathcal{C}$ denote the finite answer space defined by the task. For binary questions, $\mathcal{C} = \{\texttt{Yes}, \texttt{No}\}$; for multiple-choice questions, $\mathcal{C}$ consists of the provided answer options. For each branch $b \in \{AV, A, V\}$, let z_c^b denote the model score associated with candidate $c \in \mathcal{C}$. We define the branch prediction as

$$\hat{y}_b = \arg\max_{c \in \mathcal{C}} z_c^b, \tag{1}$$

and its decision margin as

$$m_b = z_{\hat{y}_b}^b - \max_{c \in \mathcal{C} \setminus \{\hat{y}_b\}} z_c^b. \tag{2}$$

A larger m_b indicates a more decisive prediction, while $m_b = 0$ indicates that the branch does not have a unique preferred answer. In our evaluations, z_c^b is the next-token logit assigned to the candidate answer label under the evaluation prompt.

Question-level modality routing. Inspired by the modality self-assessment mechanism of MAD (Chung et al., 2026), we use the OLLM itself to estimate which sensory modality is required by the question. We adopt a text-only routing probe: both audio and visual inputs are withheld, and the original question is followed by the prompt: "*To answer this question, which modality is needed (audio, video, or both):*" Let ρ_A, ρ_V, and ρ_B denote the logits assigned to `audio`, `video`, and `both`, respectively, and define:

$$(p_a, p_v, p_b) = \mathrm{softmax}(\rho_A, \rho_V, \rho_B). \tag{3}$$

Because the probe observes only the question, it provides a question-level modality-demand signal without accessing the audio-visual content of the current sample. The probe is used only to select a candidate counterfactual reference; it is not assumed to identify the factually correct modality. We select a modality-specific counterfactual reference only when the probe uniquely favors audio or video:

$$r = \begin{cases} A, & p_a > \max(p_v, p_b), \\ V, & p_v > \max(p_a, p_b), \\ \varnothing, & \text{otherwise.} \end{cases} \tag{4}$$

Samples routed to `both`, or for which no unique maximum exists, are assigned $r = \varnothing$. Thus, the routing signal only determines which unimodal branch, if any, serves as a candidate reference; it does not itself trigger an intervention. We evaluate the robustness of this routing signal to prompt formulation and option ordering in Section 6.3.

5 METHOD

5.1 COUNTERFACTUAL CONFLICT GATE

The routing signal identifies a candidate modality-specific reference, but disagreement with the joint audio-visual prediction is required before CaRe applies any correction. Given $r \in \{A, V, \varnothing\}$, we define the counterfactual conflict gate as

$$g = \mathbb{I}\left[r \in \{A, V\} \ \wedge \ m_r > 0 \ \wedge \ m_{AV} > 0 \ \wedge \ \hat{y}_r \neq \hat{y}_{AV}\right]. \tag{5}$$

Thus, $g = 1$ only when the routing probe provides a valid unimodal reference and both the counterfactual and joint branches make unambiguous but conflicting decisions. We refer to this disagreement as a *hard counterfactual conflict*. All other samples receive $g = 0$ and retain the original joint prediction. When $g = 1$, the selected branch r is used as the counterfactual reference for the residual intervention described in Section 5.2.

5.2 ADAPTIVE RESIDUAL INTERVENTION

The counterfactual conflict gate identifies samples for which the joint prediction disagrees with the selected modality-specific counterfactual reference, but it does not specify how the joint representation should be corrected without discarding complementary multimodal context. We therefore use the selected counterfactual branch as a representation-level reference and steer the original audio-visual inference path toward it.

For a gate-positive sample, let $h_l^r \in \mathbb{R}^d$ denote the post-residual hidden state at the final prompt-token position of decoder layer l from the selected counterfactual branch r, and let $h_l^{AV} \in \mathbb{R}^d$ denote the corresponding state on the audio-visual intervention path. We define the layer-wise counterfactual residual as

$$\Delta h_l^r = h_l^r - h_l^{AV}. \tag{6}$$

This residual gives a sample-specific direction from the current audio-visual representation toward its selected counterfactual reference. Importantly, h_l^{AV} is the current state during the intervention forward pass and therefore already reflects corrections made at preceding layers. The residual at each layer is consequently computed relative to the actual evolving audio-visual representation rather than the original unmodified state. CaRe uses this residual only to adjust the original audio-visual representation, while both modalities remain available throughout subsequent computation.

5.2.1 Confidence-Adaptive Intervention Strength

A counterfactual conflict alone does not indicate how strongly the candidate branch should influence the joint representation. In particular, a counterfactual prediction that barely separates its top candidates provides a weaker correction signal than one that makes a decisive prediction. We therefore adapt the intervention strength to the decision confidence of the selected branch.

We convert its decision margin m_r into a bounded confidence score:

$$u_r = 2\sigma(m_r) - 1 \in (0, 1), \tag{7}$$

where $\sigma(\cdot)$ denotes the sigmoid function. Larger values of u_r correspond to greater separation between the selected answer and its strongest alternative. Given a confidence threshold η, CaRe selects either a weak or strong intervention range:

$$[\alpha_{\min}, \alpha_{\max}] = \begin{cases} [\alpha_{\min}^{\mathrm{W}}, \alpha_{\max}^{\mathrm{W}}], & u_r < \eta, \\ [\alpha_{\min}^{\mathrm{S}}, \alpha_{\max}^{\mathrm{S}}], & u_r \geq \eta. \end{cases} \tag{8}$$

Low-confidence counterfactual predictions therefore receive conservative corrections, whereas more decisive predictions permit stronger intervention.

5.2.2 Depth-Decayed Residual Intervention

Applying repeated corrections with the same strength across multiple decoder layers can lead to excessive perturbation, particularly near the output. We therefore distribute the intervention over a predefined set of decoder layers while gradually reducing its strength with depth, allowing earlier corrections to propagate through subsequent computation and later layers to make milder refinements. Let

$$\mathcal{L}_{\mathrm{int}} = \{l_1, \ldots, l_K\}, \qquad l_1 < \cdots < l_K, \quad K \geq 2,$$

denote the intervention layers. Given the confidence-dependent range $[\alpha_{\min}, \alpha_{\max}]$, the coefficient at layer l_j is

$$\alpha_j = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min}) \left(\frac{K - j}{K - 1} \right)^{\gamma}, \qquad j = 1, \ldots, K, \tag{9}$$

where $\gamma > 0$ controls the decay rate. Thus, α_j decreases monotonically from $\alpha_{\max}$ at l_1 to $\alpha_{\min}$ at l_K. Following prior activation-steering work that uses the last input-token representation as a compact site for controlling subsequent generation (Stolfo et al., 2025), CaRe modifies only the final prompt-token state during prefilling. At each intervention layer l_j, we apply

$$\tilde{h}_{l_j}^{AV} = h_{l_j}^{AV} + \alpha_j \Delta h_{l_j}^r = h_{l_j}^{AV} + \alpha_j \left(h_{l_j}^r - h_{l_j}^{AV} \right). \tag{10}$$

The corrected state $\tilde{h}_{l_j}^{AV}$ is propagated through the subsequent decoder layers, so later interventions operate on the updated audio-visual representation. All other prompt-token positions remain unchanged. After prefilling, generation proceeds autoregressively from the resulting audio-visual state without further intervention on newly generated tokens.

Table 1: **Main results on CMM and AVHBench (%).** We evaluate Qwen2.5-Omni-7B and MiniCPM-o-2.6 on the Visual Dominance and Audio Dominance subsets of CMM, and on the Video-driven Audio Hallucination and Audio-driven Video Hallucination tasks of AVHBench. We report Accuracy (Acc.) and F1, with `Yes` treated as the positive class. *Overall Acc.* is computed over all samples in the two reported subsets or tasks. Best results are highlighted in bold.

Backbone	Method	CMM					AVHBench				
		Visual Dominance		Audio Dominance		Overall Acc.	Video-Driven Audio Hall.		Audio-Driven Video Hall.		Overall Acc.
		Acc.	F1	Acc.	F1		Acc.	F1	Acc.	F1	
Qwen2.5-Omni-7B	Baseline	69.3	75.7	71.8	77.0	70.5	67.6	75.3	74.1	77.4	69.8
	+VCD	75.5	79.4	66.5	73.5	71.0	68.1	75.4	77.4	78.0	71.2
	+ICT	75.3	79.1	66.5	73.6	70.9	68.8	75.6	76.7	78.4	71.4
	+AVCD	67.8	74.9	70.5	76.2	69.1	77.9	81.4	81.7	82.8	79.2
	+MAD	69.5	75.9	75.5	79.2	72.5	79.0	82.1	85.0	84.5	81.0
	+CaRe	**82.0**	**84.1**	**89.8**	**89.8**	**85.9**	**80.9**	**83.2**	**85.1**	**84.5**	**82.3**
MiniCPM-o-2.6	Baseline	78.3	81.2	67.8	73.6	73.0	72.8	74.4	74.7	75.7	73.4
	+VCD	72.3	76.3	61.0	69.5	66.6	73.9	75.0	74.0	74.5	73.9
	+ICT	72.3	76.1	62.8	70.8	67.5	72.3	73.8	75.9	76.4	73.5
	+AVCD	77.8	80.9	67.5	73.5	72.6	78.9	80.6	74.1	75.9	77.3
	+MAD	81.8	83.8	72.8	76.0	77.3	79.8	81.9	73.6	74.6	77.8
	+CaRe	**85.3**	**86.4**	**88.5**	**88.3**	**86.9**	**81.7**	**83.3**	**77.1**	**77.4**	**80.2**

6 EXPERIMENTS

6.1 EXPERIMENTAL SETUP

Models and Benchmarks. We evaluate CaRe on Qwen2.5-Omni-3B/7B (Xu et al., 2025a) and MiniCPM-o-2.6 (Yao et al., 2025), two omnimodal large language models that jointly process audio, visual, and textual inputs. We use two benchmarks for cross-modal hallucination evaluation: CMM Leng et al. (2026) and AVHBench (Sung-Bin et al., 2025). CMM examines hallucinations arising from modality over-reliance; we focus on its Visual Dominance and Audio Dominance subsets. AVHBench evaluates whether information from one modality induces false perceptions in the other; we use its Video-driven Audio Hallucination and Audio-driven Video Hallucination tasks.

Baselines and Metrics. We compare CaRe with the unmodified backbone and four inference-time hallucination-mitigation baselines: Visual Contrastive Decoding (VCD) (Leng et al., 2024), Image-Object Cross-Level Trusted Intervention (ICT) (Chen et al., 2025), Audio-Visual Contrastive Decoding (AVCD) (Jung et al., 2026), and Modality-Adaptive Decoding (MAD) (Chung et al., 2026). These methods respectively represent visual contrastive decoding, internal representation intervention, audio-visual contrastive decoding, and modality-adaptive decoding. Further details of the baseline methods are provided in Appendix C. We report accuracy (Acc.) and F1 score, with `Yes` treated as the positive class. *Overall Acc.* is computed by pooling samples across the two reported subsets or tasks within each benchmark. All values are reported in percentage.

Implementation Details. Qwen2.5-Omni-7B and MiniCPM-o-2.6 each contain 28 language-decoder layers, while Qwen2.5-Omni-3B contains 36 layers. We apply CaRe to the final three decoder layers, corresponding to layers 25–27 for Qwen2.5-Omni-7B and MiniCPM-o-2.6, and layers 33–35 for Qwen2.5-Omni-3B. We use greedy decoding with temperature 0 for deterministic evaluation. Implementation details, hyperparameter configurations, and sensitivity analyses are provided in Appendix C.

6.2 MAIN RESULTS

Table 1 reports the main results on CMM and AVHBench. CaRe achieves the best overall accuracy for both backbones on both benchmarks, and obtains the highest score in every reported subset and task. These consistent gains over contrastive decoding, representation intervention, and modality-

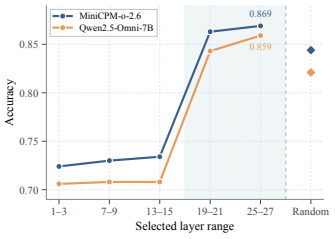

Figure 4: **Intervention-layer location.** Late layers 25–27 achieve the best accuracy across both backbones.

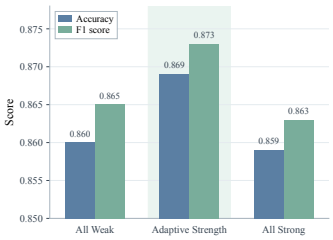

Figure 5: **Confidence-adaptive strength.** Adaptive strength outperforms uniformly weak and strong interventions.

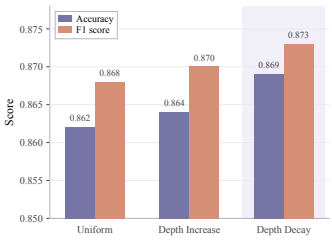

Figure 6: **Depth-decayed intervention.** Depth decay outperforms uniform and depth-increasing schedules.

adaptive decoding baselines demonstrate that CaRe effectively mitigates cross-modal hallucinations across different model architectures and failure settings.

Results on CMM. CaRe substantially improves robustness against modality over-reliance. On Qwen2.5-Omni-7B, CaRe reaches an overall accuracy of 85.9%, exceeding the strongest baseline, MAD, by 9.8 points. On MiniCPM-o-2.6, CaRe achieves 86.9%, outperforming MAD by 9.6 points. The gains are particularly pronounced on the Audio Dominance subset: CaRe improves accuracy from 75.5% to 89.8% on Qwen2.5-Omni-7B and from 72.8% to 88.5% on MiniCPM-o-2.6, relative to the strongest baseline. CaRe also consistently improves F1 scores on both dominance subsets, indicating more reliable decisions under modality-conflict scenarios.

Results on AVHBench. CaRe also achieves the best overall accuracy on both cross-modal hallucination tasks in AVHBench. For Qwen2.5-Omni-7B, CaRe attains 82.3% overall accuracy, surpassing MAD by 1.5 points. For MiniCPM-o-2.6, it reaches 80.2%, improving over the strongest baseline, MAD, by 2.5 points. CaRe obtains the highest accuracy and F1 score on both Video-driven Audio Hallucination and Audio-driven Video Hallucination tasks for both backbones. These results show that the proposed conflict-aware residual intervention generalizes to hallucinations caused by interference in either modality direction.

Robustness across model scales. To examine whether CaRe depends on model capacity, we additionally evaluate Qwen2.5-Omni-3B, providing a controlled comparison with its 7B counterpart within the same model family. On the 3B model, CaRe improves overall accuracy from 76.9% to 82.5% on CMM and from 76.4% to 81.0% on AVHBench, yielding gains of 5.6 and 4.6 points, respectively. Together with the improvements observed on Qwen2.5-Omni-7B, these results show that CaRe remains effective across model scales rather than being specific to a single capacity setting. Detailed category-wise results are provided in Appendix B.

Table 2: Scaling robustness on Qwen2.5-Omni models. Gains over Base are shown in parentheses.

Backbone	Method	CMM	AVHBench
		Acc.	Acc.
Qwen2.5-Omni-3B	Base	76.9	76.4
	+CaRe	**82.5** (+5.6)	**81.0** (+4.6)
Qwen2.5-Omni-7B	Base	70.5	69.8
	+CaRe	**85.9** (+15.4)	**82.3** (+12.5)

6.3 Ablation Studies

We conduct component-wise ablations to examine the main design choices of CaRe. Unless otherwise specified, all experiments are conducted on CMM and aggregate results over the Visual and Audio Dominance subsets. Each ablation varies one component while keeping the remaining settings fixed.

Residual intervention vs. branch selection. To isolate the benefit of representation-level correction, we compare CaRe with CF-SELECT, which directly outputs the counterfactual prediction $\hat{y}^r$ when a conflict is detected ($g = 1$). Since both methods behave identically on gate-negative samples, we evaluate directly on the gate-positive subset where their decisions diverge. As shown in Table 3,

Table 3: **Residual intervention vs. direct branch selection on gate-positive samples (%).** Evaluations are conducted exclusively on conflict-gated samples ($g = 1$). CF-SELECT directly outputs the counterfactual unimodal prediction $\hat{y}^r$, whereas CaRe applies continuous residual steering to the joint representation. Accuracy gains of CaRe over CF-SELECT are reported in parentheses at the bottom right.

Backbone	Method	CMM		AVHBench	
		Visual Dom.	Audio Dom.	V→A Hall.	A→V Hall.
Qwen2.5-Omni-7B	CF-SELECT	80.14	86.46	54.22	53.61
	+CaRe	**81.69** (+1.55)	**88.54** (+2.08)	**63.05** (+8.84)	**62.89** (+9.28)
MiniCPM-o-2.6	CF-SELECT	66.67	84.43	48.56	60.38
	+CaRe	**73.02** (+6.35)	**86.49** (+2.06)	**59.38** (+10.82)	**64.15** (+3.77)

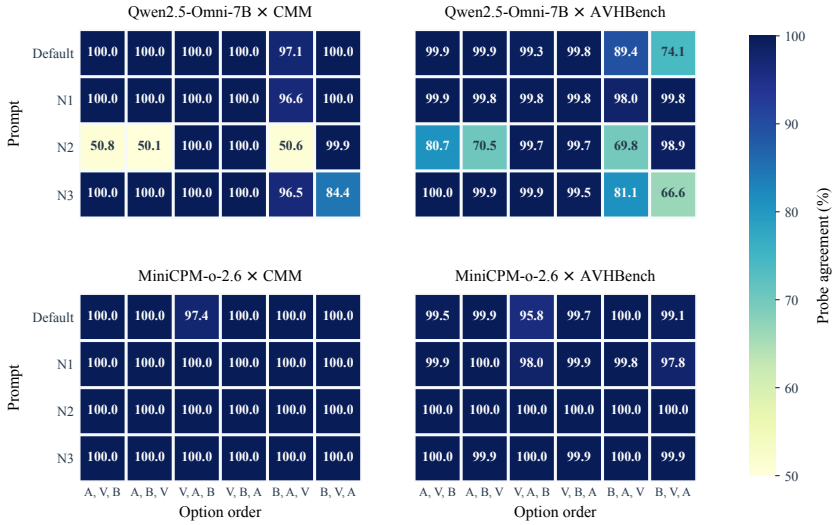

Figure 7: **Robustness of text-only modality routing.** Probe agreement (%) with the task-defined target modality across prompt formulations and all six modality-label orderings. Results are shown for both backbones on CMM and AVHBench; prompt templates are provided in Appendix D.

CaRe consistently outperforms CF-SELECT across both benchmarks and backbones, achieving accuracy gains of up to 10.82 points on AVHBench and 6.35 points on CMM. This confirms that the advantage of CaRe stems from steering the multimodal representation rather than trivially switching to a unimodal branch: by using the counterfactual signal as a continuous residual direction, CaRe resolves modality misrouting while retaining complementary cross-modal context.

Robustness of text-only modality routing. Since CaRe relies on the text-only probe to select a counterfactual reference, we examine the robustness of its routing decisions to prompt formulation and option ordering. For CMM and AVHBench, the question-relevant modality is specified by task construction, allowing us to measure probe agreement with the task-defined target modality. As shown in Figure 7, we evaluate four prompt formulations and all six permutations of `audio`, `video`, and `both` across the four backbone–benchmark settings. Among the resulting 96 configurations, 85 achieve at least 95% agreement, while the default configuration used in all main experiments achieves 99.5–100% agreement across the four settings. Overall, the routing signal is broadly stable, with some model-specific sensitivity to prompt and option ordering. The exact probe formulations are provided in Appendix D.

Intervention-layer location. Figure 4 compares different three-layer intervention ranges. Interventions in early and intermediate layers provide limited benefits, whereas performance improves sharply in later layers. Compared with layers 19–21, the default range of layers 25–27 further improves accuracy by 1.6 points on Qwen2.5-Omni-7B and 0.6 points on MiniCPM-o-2.6. It also outperforms random layer selection by 3.8 and 2.5 points, respectively. These results support applying residual correction near the output stage, where the fused representation is more directly associated with the final decision.

Table 4: **Accuracy on Daily-Omni (%).** Results are reported across question categories and video durations. The better result within each backbone is highlighted in bold.

Backbone	Method	All	Question Category						Duration	
			Event Seq.	AV Event Align.	Inference	Reasoning	Context Understanding	Comparative	30s	60s
Qwen2.5-Omni-7B	Base	61.5	55.6	50.4	75.3	75.4	**57.5**	**66.4**	62.6	60.2
	CaRe	**61.8**	**56.2**	**51.3**	**76.6**	**76.6**	56.5	64.9	**62.9**	**60.5**
MiniCPM-o-2.6	Base	61.0	55.6	**48.7**	76.0	72.6	**60.6**	63.4	61.1	60.9
	CaRe	**61.5**	**55.9**	**48.7**	**77.3**	**73.7**	**60.6**	**64.1**	**61.8**	**61.1**

Confidence-adaptive intervention strength. Figure 5 compares adaptive strength with uniformly weak and uniformly strong interventions on CMM using MiniCPM-o-2.6. Adaptive strength improves accuracy and F1 by 0.9 and 0.8 points over the all-weak setting, and by 1.0 point on both metrics over the all-strong setting. This suggests that weak correction can be insufficient for decisive counterfactual predictions, whereas uniformly strong correction may over-perturb less certain cases. Adapting the intervention strength to counterfactual confidence provides a better balance.

Depth-decayed residual intervention. Figure 6 compares the proposed depth-decayed schedule with uniform strength and depth-increasing strength on CMM using MiniCPM-o-2.6. The depth-decayed schedule performs best, achieving 0.869 accuracy and 0.873 F1, compared with 0.862/0.868 for uniform strength and 0.864/0.870 for depth-increasing strength. The gain does not arise merely from varying the intervention coefficients across layers; assigning stronger corrections to earlier layers in the intervention range and weaker corrections to deeper layers is particularly effective.

6.4 Performance on General Audio-Visual QA Tasks

To assess CaRe beyond hallucination-focused benchmarks, we evaluate it on Daily-Omni (Zhou et al., 2025), a general audio-visual QA benchmark covering event sequencing, audio-visual alignment, inference, reasoning, contextual understanding, and comparative analysis. As shown in Table 4, CaRe increases overall accuracy from 61.5% to 61.8% on Qwen2.5-Omni-7B and from 61.0% to 61.5% on MiniCPM-o-2.6. The most consistent category-level gains are observed in inference and reasoning, with improvements of 1.3 points on inference for both backbones and 1.2/1.1 points on reasoning. CaRe also yields slightly higher accuracy on both 30-second and 60-second videos. Overall, these results indicate that applying CaRe does not compromise aggregate performance on general audio-visual QA, while yielding small overall accuracy gains on both backbones.

7 Conclusion

Our results show that some audio-visual hallucinations occur even when task-relevant evidence remains recoverable from a modality-specific view, motivating modality misrouting as a useful characterization of such failures. Based on this observation, we introduce CaRe, which detects conflicts between the joint prediction and a question-relevant counterfactual view, and uses the resulting residual signal to correct fused representations. Its confidence-adaptive and depth-decayed intervention enables targeted correction at middle-to-late decoder layers without updating model parameters. Experiments on CMM and AVHBench demonstrate substantial improvements across two OLLMs, while Daily-Omni results show that CaRe also preserves and modestly improves general audio-visual question-answering performance.

Limitations and Future Work. CaRe currently assumes that useful task-relevant evidence remains accessible in a modality-specific branch, and its conflict gate is formulated over a discrete task-defined answer space where branch predictions can be directly compared. As a result, it may be less effective when relevant evidence is weakly encoded, when modality misrouting does not manifest as an explicit prediction conflict, or when the task does not provide a readily scorable candidate answer space. In addition, the text-only routing probe and counterfactual branches introduce extra inference overhead. Future work may explore softer conflict signals, more general routing criteria for open-ended generation, and more efficient single-pass implementations.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, improving the clarity and readability of the manuscript, and organizing and formatting data. They were not used to formulate the research methodology, generate or interpret experimental results, or derive the scientific conclusions of this work. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We provide detailed information to facilitate reproduction of our results. Sections 5.1–5.2 describe the complete formulation of CaRe, and Section 6.1 specifies the evaluated models, benchmarks, metrics, and experimental protocol. Appendix A provides additional ablations and hyperparameter analyses, including the counterfactual conflict gate, intervention strength, confidence threshold, and their joint sensitivity. Appendix C provides the complete CaRe inference algorithm, hyperparameter configurations, selection protocol, and implementation details for all baselines. Appendix D reports the exact prompt formulations used for the text-only modality probe, while Appendix E provides additional details on the evaluation datasets. Finally, Appendix F reports end-to-end inference cost, runtime decomposition, and the potential latency reduction under ideal branch-level parallel execution.

REFERENCES

Ami Baid, Zihui Xue, and Kristen Grauman. Don't let the video speak: Audio-contrastive preference optimization for audio-visual language models. *arXiv preprint arXiv:2604.14129*, 2026.

Ashutosh Chaubey, Jiacheng Pang, and Mohammad Soleymani. Mod-dpo: Towards mitigating cross-modal hallucinations in omni llms using modality decoupled preference optimization. *arXiv preprint arXiv:2603.03192*, 2026.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. Ict: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4209–4221. IEEE, 2025.

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

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.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction, 2026. URL https://arxiv.org/abs/2604.27393.

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)*, pp. 14303–14312. IEEE, 2024.

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

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

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

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27026–27036. IEEE, 2024.

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

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

Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. *Advances in Neural Information Processing Systems*, 38, 2026.

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, 2023.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 35799–35819. PMLR, 2025.

Run Luo, Ting-En Lin, Haonan Zhang, Yuchuan Wu, Xiong Liu, Yongbin Li, Longze Chen, Jiaming Li, Lei Zhang, Xiaobo Xia, et al. Openomni: Advancing open-source omnimodal large language models with progressive multimodal alignment and real-time emotional speech synthesis. *Advances in Neural Information Processing Systems*, 38:158925–158953, 2026.

Ji Ma, Wei Suo, Peng Wang, and Yanning Zhang. Understanding and mitigating hallucinations in multimodal chain-of-thought models. *arXiv preprint arXiv:2603.27201*, 2026.

Miao Pan, Wangjie Gan, Jintao Chen, Wenqi Zhang, Sun Bing, Jianwei Yin, and Xuhong Zhang. Ground what you see: Hallucination-resistant mllms via caption feedback, diversity-aware sampling, and conflict regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8242–8250, 2026.

Jean Park, Kuk Jin Jang, Basam Alasaly, Sriharsha Mopidevi, Andrew Zolensky, Eric Eaton, Insup Lee, and Kevin Johnson. Assessing modality bias in video question answering benchmarks with multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19821–19829, 2025.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering. In *International Conference on Learning Representations*, volume 2025, pp. 55790–55823, 2025.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12964–12974. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.634.

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.

Qwen Team. Qwen3.5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Junfei Wu, Yue Ding, Guofan Liu, Tianze Xia, Ziyue Huang, Dianbo Sui, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. SHARP: Steering hallucination in LVLMs via representation engineering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14346–14361. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.725.

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

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

Jianjiang Yang, Peihang Li, Shanqing Xu, Mengchen Qian, Lu Zhang, and Meng Luo. Omnihallu: Unified hallucination detection for cross-modal comprehension and generation in multimodal large language models, 2026. URL https://arxiv.org/abs/2609.11244.

Yuan Yao, Tianyu Yu, Chongyi Wang, Junbo Cui, Bokai Xu, Hongji Zhu, Tianchi Cai, Fuwei Huang, Tianran Wang, Wenshuo Ma, et al. Minicpm-o 2.6: A gpt-4o level mllm for vision, speech, and multimodal live streaming on your phone, 2025.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Zhen Wan, Jinchuan Tian, An-Chieh Cheng, Ligeng Zhu, Yuanhang Su, Yuming Lou, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. In *International Conference on Learning Representations*, volume 2026, pp. 56101–56138, 2026.

Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. Dynamic multimodal activation steering for hallucination mitigation in large vision-language models. In *International Conference on Learning Representations*, 2026.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024.

Bai Zechen et al. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

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

Zongmeng Zhang, Wengang Zhou, Jie Zhao, and Houqiang Li. Robust multimodal large language models against modality conflict. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 77233–77253. PMLR, 2025.

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

APPENDIX OVERVIEW

Appendix	Title
Appendix A	**Additional Ablation and Hyperparameter Analysis**
↪ Appendix A.1	Ablation of the Counterfactual Conflict Gate
↪ Appendix A.2	Impact of the Intervention Strength Scale λ_α
↪ Appendix A.3	Impact of the Confidence Threshold η
↪ Appendix A.4	Joint Sensitivity to λ_α and η
Appendix B	**Scaling Analysis**
Appendix C	**Reproducibility Details**
↪ Appendix C.1	CaRe Inference Algorithm
↪ Appendix C.2	Hyperparameter Configurations
↪ Appendix C.3	Baseline Implementation Details
Appendix D	**Prompt Formulations for the Text-Only Modality Probe**
Appendix E	**Additional Dataset Details**
↪ Appendix E.1	CMM
↪ Appendix E.2	AVHBench
↪ Appendix E.3	Daily-Omni
Appendix F	**Computational Analysis**

A ADDITIONAL ABLATION AND HYPERPARAMETER ANALYSIS

A.1 ABLATION OF THE COUNTERFACTUAL CONFLICT GATE

We compare the full CaRe gate with a *routing-only* variant that applies residual correction whenever the probe identifies a unique modality-specific candidate, without requiring a counterfactual conflict. As shown in Table 5, removing the conflict condition increases intervention coverage by approximately 75 percentage points on both backbones, yet leaves accuracy unchanged. Moreover, none of these additional interventions changes prediction correctness. This result indicates that the conflict gate effectively excludes redundant corrections and concentrates intervention on samples with an explicit disagreement between the joint and counterfactual decisions.

Table 5: **Ablation of the counterfactual conflict gate on CMM.** *Routing-only* applies residual correction whenever the probe returns a unique unimodal candidate. *Help/Harm* counts additional routing-only interventions that change an originally incorrect/correct prediction into a correct/incorrect one, respectively.

Backbone	Intervention Coverage		Accuracy	Extra Interventions (Help / Harm)
	Conflict-gated	Routing-only	Conflict / Routing	
Qwen2.5-Omni-7B	167/800 (20.9%)	773/800 (96.6%)	85.9 / 85.9	606 (0 / 0)
MiniCPM-o-2.6	174/800 (21.8%)	772/800 (96.5%)	86.9 / 86.9	598 (0 / 0)

A.2 IMPACT OF THE INTERVENTION STRENGTH SCALE λ_α

We examine the sensitivity of CaRe to the overall intervention magnitude on CMM. Specifically, after confidence adaptation and depth decay determine the layer-wise coefficient α_j, we rescale it as $\widetilde{\alpha}_j = \lambda_\alpha \alpha_j$, while keeping all other settings fixed. Figure 8 reports both the resulting accuracy and the numbers of helpful and harmful prediction changes relative to the unmodified backbone.

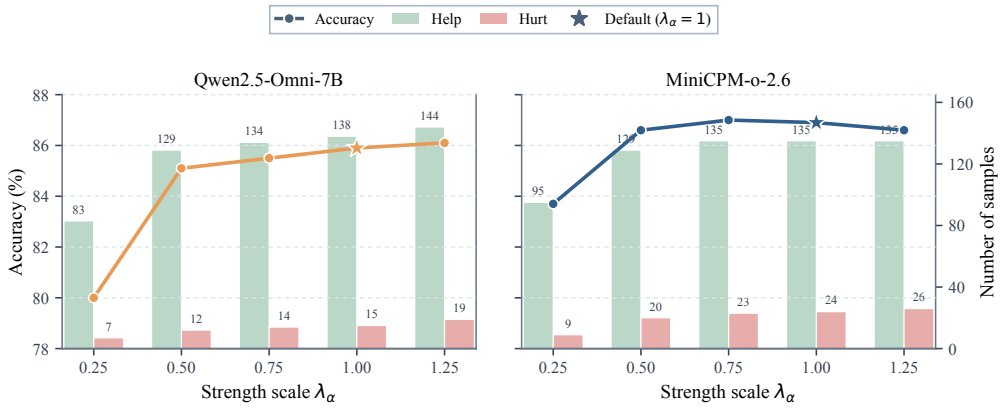

Figure 8: **Sensitivity to the intervention strength scale λ_α on CMM.** Lines report accuracy, while bars show the numbers of helpful and harmful prediction changes. Stars denote the default setting $\lambda_\alpha = 1$.

Performance improves substantially when λ_α increases from 0.25 to 0.50, and remains stable over the moderate range from 0.50 to 1.25. Increasing the scale generally produces more helpful corrections, but also introduces more harmful changes, indicating a trade-off between correction capability and over-intervention. The default $\lambda_\alpha = 1$ achieves near-optimal accuracy on both backbones, falling only 0.2 and 0.1 percentage points below their respective best settings. This result suggests that CaRe is not sensitive to a narrowly tuned intervention scale within the effective range.

A.3 Impact of the Confidence Threshold η

We further examine the sensitivity of CaRe to the confidence threshold η, which determines whether a gate-positive sample receives a weak or strong residual intervention. Specifically, samples with counterfactual confidence $u_r < \eta$ are assigned to the weak regime, while those with $u_r \geq \eta$ receive the strong intervention. We vary η from 0.1 to 0.9 while keeping all other hyperparameters fixed. Figure 9 reports the resulting accuracy together with the numbers of samples assigned to the weak and strong intervention regimes.

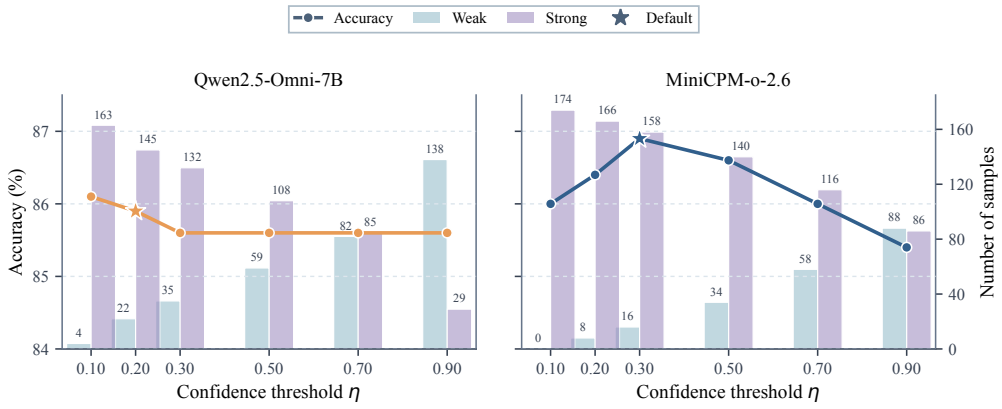

Figure 9: **Sensitivity to the confidence threshold η on CMM.** Lines report accuracy, while bars show the numbers of samples assigned to the weak and strong intervention regimes. Stars denote the default settings, $\eta = 0.2$ for Qwen2.5-Omni-7B and $\eta = 0.3$ for MiniCPM-o-2.6.

As η increases, the intervention allocation shifts monotonically from the strong regime toward the weak regime. For Qwen2.5-Omni-7B, the number of weak interventions increases from 4 at $\eta = 0.1$ to 138 at $\eta = 0.9$, while strong interventions decrease from 163 to 29. A similar trend is observed

for MiniCPM-o-2.6, where the weak/strong split changes from $0/174$ to $88/86$. This confirms that η directly controls the conservativeness of CaRe: a lower threshold applies strong correction to most conflict-positive samples, whereas a higher threshold reserves strong intervention for only the most confident counterfactual predictions.

The two backbones exhibit different sensitivity patterns. Qwen2.5-Omni-7B achieves its highest accuracy of 86.1% at $\eta = 0.1$, while the default $\eta = 0.2$ attains 85.9%, only 0.2 percentage points lower. Performance becomes nearly unchanged at 85.6% for larger thresholds. In contrast, MiniCPM-o-2.6 exhibits a clearer intermediate optimum: accuracy increases from 86.0% at $\eta = 0.1$ to 86.9% at the default $\eta = 0.3$, and then gradually decreases as the threshold is further increased, reaching approximately 85.4% at $\eta = 0.9$. These results suggest that overly aggressive strong intervention can be suboptimal when counterfactual confidence is insufficient, whereas an excessively large threshold makes the correction overly conservative and prevents sufficiently confident samples from receiving stronger intervention. Overall, a moderate, backbone-specific confidence threshold provides an effective balance between correction strength and intervention reliability.

A.4 Joint Sensitivity to λ_α and η

While λ_α controls the overall magnitude of residual correction, the confidence threshold η determines how frequently the strong intervention regime is activated. We therefore evaluate their joint effect on CMM using a 3×3 grid for each backbone. Figure 10 reports the resulting accuracy.

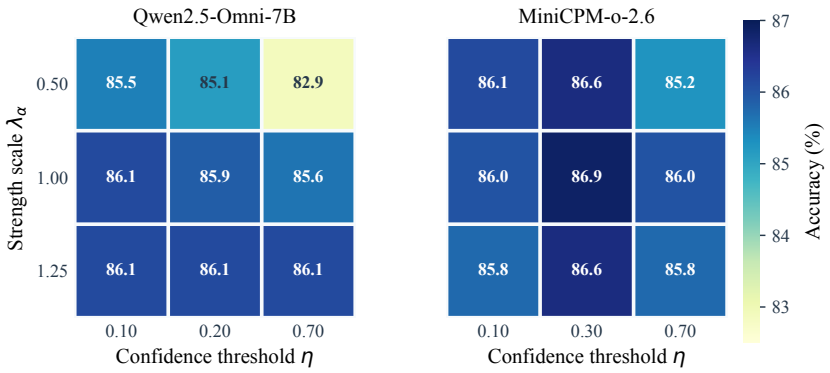

Figure 10: **Joint sensitivity to the intervention strength scale λ_α and confidence threshold η on CMM.** Each cell reports accuracy (%); darker colors indicate higher accuracy.

The two hyperparameters exhibit a clear interaction. For Qwen2.5-Omni-7B, a small strength scale becomes sensitive to the threshold: performance drops to 82.875% when $\lambda_\alpha = 0.5$ and $\eta = 0.7$. Increasing λ_α mitigates this under-correction, and $\lambda_\alpha = 1.25$ maintains 86.125% accuracy across all evaluated thresholds. For MiniCPM-o-2.6, the best result of 86.875% is obtained at the intermediate configuration $(\lambda_\alpha, \eta) = (1.0, 0.3)$. Increasing the scale to 1.25 does not yield further gains and slightly degrades performance away from this confidence threshold.

Overall, the results show that intervention magnitude and confidence-based allocation should be calibrated jointly rather than independently. The default configurations lie in high-performing regions for both backbones.

B Scaling Analysis

To further examine whether CaRe generalizes across model capacities, we evaluate Qwen2.5-Omni-3B and Qwen2.5-Omni-7B under the same evaluation protocol. Table 6 reports detailed results on each subset of CMM and AVHBench. CaRe consistently improves overall performance while maintaining gains across different cross-modal hallucination directions.

Table 6: **Detailed scaling results on Qwen2.5-Omni-3B (%).** We compare the base model with CaRe on CMM and AVHBench. Results are reported as Accuracy (Acc.) and F1, with `Yes` treated as the positive class. *Overall Acc.* is computed over all samples in the corresponding subsets or tasks. Best results are highlighted in bold.

Backbone	Method	CMM					AVHBench				
		Visual Dominance		Audio Dominance		Overall Acc.	Video-Driven Audio Hall.		Audio-Driven Video Hall.		Overall Acc.
		Acc.	F1	Acc.	F1		Acc.	F1	Acc.	F1	
Qwen2.5-Omni-3B	Base	69.3	66.5	84.5	82.5	76.9	74.2	74.5	80.8	77.5	76.4
	+CaRe	**79.8**	**79.1**	**85.3**	**83.5**	**82.5**	**81.0**	**81.4**	**80.9**	**77.6**	**81.0**

Table 7: Hyperparameter configurations of CaRe on CMM and AVHBench.

Backbone	Dataset	Intervention Layers	γ	η	α^W	α^S
Qwen2.5-Omni-7B	CMM	25–27	0.5	0.2	$[0.2, 0.5]$	$[0.5, 0.8]$
	AVHBench	25–27	0.5	0.2	$[0.2, 0.5]$	$[0.5, 0.8]$
Qwen2.5-Omni-3B	CMM	33–35	0.5	0.2	$[0.2, 0.5]$	$[0.5, 0.8]$
	AVHBench	33–35	0.5	0.2	$[0.2, 0.5]$	$[0.5, 0.8]$
MiniCPM-o-2.6	CMM	25–27	0.5	0.3	$[0.1, 0.3]$	$[0.5, 0.8]$
	AVHBench	25–27	0.5	0.5	$[0.1, 0.3]$	$[0.3, 0.5]$

C Reproducibility Details

C.1 CaRe Inference Algorithm

Algorithm 1 summarizes the complete inference procedure of CaRe. Given an audio input a, a visual input v, and a textual question q, CaRe first performs the original audio-visual inference and uses a text-only modality-demand probe to identify a candidate question-relevant modality. A modality-specific counterfactual branch is constructed only when the probe uniquely favors audio or video. CaRe then compares the prediction of the candidate branch with that of the original joint branch and activates residual correction only when the two branches exhibit a hard decision conflict. Otherwise, the original audio-visual output is retained without modification.

For a gate-positive sample, CaRe computes layer-wise counterfactual residuals between the candidate unimodal and joint audio-visual representations. The decision margin of the candidate branch is converted into a bounded confidence score, which determines whether a weak or strong intervention range is used. Within the predefined intervention layers, the intervention strength decreases with depth, and the residual correction is applied only to the hidden state of the final prompt token. After this prompt-level correction, inference continues along the original audio-visual path with both modalities preserved, and the complete response is generated autoregressively without further intervention.

C.2 Hyperparameter Configurations

Table 7 summarizes the hyperparameter configurations used for CaRe across the evaluated settings. Qwen2.5-Omni-7B and MiniCPM-o-2.6 use layers 25–27, while Qwen2.5-Omni-3B uses layers 33–35, corresponding to the final three language-decoder layers of each backbone. We further examine the sensitivity of the main hyperparameters in Appendix A.2 and Appendix A.3.

C.3 Baseline Implementation Details

For ICT (Chen et al., 2025), AVCD (Jung et al., 2026), and MAD (Chung et al., 2026), we strictly adopt their official open-source implementations, directly interfacing them with the omnimodal backbone models.

Algorithm 1 CaRe Inference

Require: Frozen model f_θ; audio a; video v; question q; intervention layers $\mathcal{L}_{\text{int}} = \{l_1, \ldots, l_K\}$; threshold η; weak/strong ranges $[\alpha^W_{\min}, \alpha^W_{\max}]$ and $[\alpha^S_{\min}, \alpha^S_{\max}]$; decay exponent γ.
Ensure: CaRe output.

```
Run the joint branch B_AV = (a, v, q) to obtain ŷ_AV and m_AV.
Run the text-only modality probe to obtain (p_a, p_v, p_b).
if p_a > max(p_v, p_b) then
    r ← A, B_r ← (a, q).
else if p_v > max(p_a, p_b) then
    r ← V, B_r ← (v, q).
else
    return original audio-visual output.
end if
Run B_r to obtain ŷ_r, m_r, and {h^r_{l_j}}_{j=1}^K.
if m_r = 0 or m_AV = 0 or ŷ_r = ŷ_AV then
    return original audio-visual output.
end if
u_r ← 2σ(m_r) − 1.
[α_min, α_max] ← { [α^W_min, α^W_max],  u_r < η,
                  { [α^S_min, α^S_max],  u_r ≥ η.
Re-run B_AV for prompt prefilling with interventions at L_int.
for j = 1, ..., K do
    At layer l_j, let h^AV_{l_j} denote the current post-residual hidden state at the final prompt-token position.
    Δh^r_{l_j} ← h^r_{l_j} − h^AV_{l_j}.
    α_j ← α_min + (α_max − α_min) ((K − j)/(K − 1))^γ.
    h̃^AV_{l_j} ← h^AV_{l_j} + α_j Δh^r_{l_j}.
    Propagate h̃^AV_{l_j} through the subsequent decoder layers.
end for
Continue autoregressive generation from the intervened audio-visual prefilling state without further intervention.
return CaRe output.
```

For Visual Contrastive Decoding (VCD) (Leng et al., 2024), the original method was designed solely for vision-language models by contrasting original representations against visual perturbations. To adapt VCD to the omnimodal setting in a principled and symmetric manner, we implement VCD by contrasting the complete audio-visual input against all possible unimodal and text-only sub-branches, withholding competing sensory inputs rather than injecting heuristic noise. Formally, let v and a represent the visual and audio inputs, and let $\emptyset$ denote a withheld modality. At each decoding step t, the calibrated logit $\mathbf{z}_t^{\text{ext}}$ is computed as:

$$\mathbf{z}_t^{\text{ext}} = \mathbf{z}_t(v, a, q) + \alpha \sum_{(v', a') \in \mathcal{S}} \left(\mathbf{z}_t(v, a, q) - \mathbf{z}_t(v', a', q) \right), \tag{11}$$

where $\mathcal{S} = \{(v, \emptyset), (\emptyset, a), (\emptyset, \emptyset)\}$ collects the counterfactual input configurations obtained by withholding the audio input, visual input, or both, respectively.

D Prompt Formulations for the Text-Only Modality Probe

We provide the exact prompt formulations used in the modality-routing robustness analysis. For all variants, both audio and visual inputs are withheld, and the routing prompt is appended to the original textual question. The placeholder {`options`} is instantiated with an ordered list of `audio`, `video`, and `both`. The main experiments use prompt **D** with the ordering `audio, video, both`, while the robustness analysis in Section 6.3 additionally considers alternative prompt formulations and option orderings.

Table 8: **Prompt formulations for the text-only modality probe. D** is used throughout the main experiments, while **N1–N3** are alternative formulations considered in the robustness analysis.

ID	Prompt template
D	To answer this question, which modality is needed ({options}):
N1	Determine the minimum evidence needed to answer the question. Select one modality: {options}:
N2	What kind of input is necessary for a correct answer? Choose from: {options}:
N3	Classify the question by the information source it requires. Answer with one of: {options}:

Table 8 lists the default prompt **D** and three alternative formulations, **N1–N3**, which share the same modality-routing objective.

E Additional Dataset Details

E.1 CMM

CMM (Leng et al., 2026) is designed to evaluate hallucinations in multimodal language models arising from over-reliance on unimodal priors and spurious inter-modality correlations. The benchmark contains six subcategories, among which we focus on *Visual Dominance* and *Audio Dominance* to study bidirectional audio-visual interference. In Visual Dominance, salient visual content may cause the model to hallucinate an associated but absent sound event, whereas in Audio Dominance, auditory cues may induce hallucinated visual objects or events. Each subcategory contains 200 multimodal instances, with each instance paired with one positive question about an existent object/event and one negative question about a non-existent one, yielding 400 binary Yes–No questions per subcategory. We therefore evaluate on 800 questions in total across the two dominance subsets.

E.2 AVHBench

AVHBench (Sung-Bin et al., 2025) is an audio-visual hallucination benchmark containing 2,136 videos and four tasks covering cross-modal judgment and description. We use its two cross-modal hallucination tasks: *Audio-driven Video Hallucination* and *Video-driven Audio Hallucination*. The former evaluates whether auditory evidence causes a model to falsely perceive an object or event as visually present, while the latter examines whether visible content induces a hallucinated sound event. Both tasks are formulated as balanced binary Yes–No questions. Audio-driven Video Hallucination contains 1,136 questions (568 Yes and 568 No), while Video-driven Audio Hallucination contains 2,290 questions (1,145 Yes and 1,145 No). These complementary directions allow us to evaluate whether CaRe mitigates cross-modal interference when either audio or visual information acts as the misleading modality.

E.3 Daily-Omni

Daily-Omni (Zhou et al., 2025) is a general audio-visual question-answering benchmark designed to evaluate joint perception and reasoning in daily-life scenarios. It contains 684 videos and 1,197 multiple-choice QA pairs spanning six task categories: *Event Sequence*, *Audio-Visual Event Alignment*, *Inference*, *Reasoning*, *Context Understanding*, and *Comparative Analysis*. The videos are segmented into 30-second and 60-second clips, corresponding to 647 and 550 questions, respectively, enabling evaluation across different temporal contexts. Unlike CMM and AVHBench, which explicitly target hallucination under cross-modal interference, Daily-Omni requires broader integration of audio, visual, and textual information. We use the full benchmark to evaluate whether CaRe preserves general audio-visual understanding beyond hallucination-focused settings, and report both category-wise and duration-wise accuracy.

Table 9: **End-to-end inference latency under serial execution.** Latency is measured over 200 samples for each backbone. We report the mean, standard deviation, median, and 95th-percentile (P95) latency in seconds.

Backbone	Method	Mean	Std.	Median	P95
Qwen2.5-Omni-7B	CaRe	6.698	4.965	5.438	15.934
	MAD	16.612	25.150	8.797	50.130
	AVCD	4.838	2.852	4.262	11.022
MiniCPM-o-2.6	CaRe	4.571	4.053	3.416	12.075
	MAD	7.816	4.157	6.245	16.573
	AVCD	2.966	1.973	2.142	7.636

Table 10: **Average per-sample latency breakdown of CaRe.** All values are reported in seconds and averaged over the same 200-example evaluation subset. *Counterfactual* denotes inference with the selected modality-specific branch, and *Residual Corr.* denotes the live residual-intervention stage.

Backbone	Probe	Joint AV	Counterfactual	Gate/Control	Residual Corr.
Qwen2.5-Omni-7B	0.021	4.369	2.004	< 0.001	0.305
MiniCPM-o-2.6	0.021	3.031	1.479	< 0.001	0.040

F Computational Analysis

We analyze the inference-time computational cost of CaRe and compare it with AVCD and MAD. All latency measurements are obtained on the same 200-example evaluation subset for each backbone under serial execution. We first report the measured end-to-end latency, then decompose the runtime of CaRe into its major stages. Finally, we estimate the potential latency reduction under ideal branch-level parallel execution.

End-to-end latency. Table 9 reports the end-to-end wall-clock latency under serial execution. AVCD achieves the lowest mean latency on both backbones, while CaRe remains substantially faster than MAD. On Qwen2.5-Omni-7B, CaRe requires 6.70 s per sample on average, compared with 16.61 s for MAD, corresponding to a 59.7% reduction. On MiniCPM-o-2.6, the mean latency decreases from 7.82 s with MAD to 4.57 s with CaRe, a reduction of 41.5%. The additional cost of CaRe relative to AVCD arises primarily from the modality-specific counterfactual inference required for conflict diagnosis.

Runtime decomposition of CaRe. To identify the source of CaRe's computational overhead, Table 10 decomposes its average per-sample latency into the text-only modality probe, the original audio-visual inference, the selected modality-specific counterfactual inference, the gating and control logic, and the residual-correction stage. The original audio-visual and counterfactual branches account for most of the runtime, whereas the text-only probe and conflict-gating logic introduce negligible overhead. The residual-correction stage also contributes only a small fraction of the total latency. In particular, its average cost is 0.305 s on Qwen2.5-Omni-7B and 0.040 s on MiniCPM-o-2.6.

Table 11: **Theoretical latency under ideal branch-level parallel execution.** Serial latency is measured directly, while parallel latency is estimated from the measured stage-wise runtimes assuming independent branches can execute concurrently without scheduling, communication, or device-synchronization overhead. Reduction is relative to serial execution.

Backbone	Method	Serial	Parallel	Reduction
Qwen2.5-Omni-7B	CaRe	6.698	4.782	28.61%
	MAD	16.612	12.038	27.53%
	AVCD	4.838	4.268	11.80%
MiniCPM-o-2.6	CaRe	4.571	3.239	29.15%
	MAD	7.816	5.374	31.25%
	AVCD	2.966	2.910	1.89%

Potential for parallel execution. A substantial portion of the serial cost comes from inference branches that need not be executed strictly sequentially. In CaRe, the modality probe is lightweight and determines which modality-specific counterfactual branch is required. Once the candidate modality is identified, the selected counterfactual computation can overlap with the original audio-visual inference, providing natural branch-level parallelism before the conflict decision and residual correction. We therefore estimate the theoretical latency obtainable under ideal parallel execution using the measured stage-wise runtimes.

As shown in Table 11, the estimated latency of CaRe decreases from 6.70 s to 4.78 s on Qwen2.5-Omni-7B and from 4.57 s to 3.24 s on MiniCPM-o-2.6, corresponding to reductions of 28.6% and 29.2%, respectively. This suggests that a substantial fraction of CaRe's serial overhead originates from independently executable diagnostic computation, rather than from the residual intervention itself. MAD exhibits a similar potential benefit from parallel branch execution, whereas the estimated reduction for AVCD is smaller.