Diagnosing and Mitigating Modality Interference in Multimodal Large Language Models

Rui Cai **Bangzheng Li** **Xiaofei Wen** **Muhao Chen** **Zhe Zhao**
University of California, Davis
ruicai@ucdavis.edu

Github: https://github.com/luisrui/Modality-Interference-in-MLLMs.git

Abstract

Multimodal Large Language Models demonstrate strong performance on multimodal benchmarks, yet remain fragile when one modality carries spurious or misleading content. We trace this fragility to a deficiency in cross-modality competency, defined as the ability to fairly evaluate and integrate information across modalities, and identify its concrete, measurable manifestation as *Modality Interference*, in which task-irrelevant modality signals improperly influence the model's predictions. To diagnose this phenomenon systematically, we design a perturbation-based evaluation grounded in causal intervention, in which controlled noise is injected into the task-irrelevant modality across image-heavy and text-heavy tasks. Across diverse MLLM families and scales from 2B to 110B parameters, we observe consistent performance degradation when perturbing the irrelevant modality, indicating that modality interference is a pervasive failure mode rather than an artifact of any specific architecture. To mitigate this, we propose a unified perturbation-aware fine-tuning framework that combines (i) heuristic and adversarial data augmentation targeting the task-irrelevant modality, and (ii) output-level consistency regularization between clean and perturbed inputs. Extensive experiments across diverse MLLM architectures, model scales, and benchmarks show that our approach simultaneously improves unimodal robustness and standard multimodal performance, achieving Pareto-optimal gains over existing baselines.

1 Introduction

Multimodal Large Language Models (MLLMs) have made significant strides in integrating vision and language understanding within a unified architecture [1, 2, 3]. By combining powerful visual encoders and large language models through alignment mechanisms, MLLMs such as LLaVA [1] and Qwen-VL [3] demonstrate strong capabilities across a wide range of multimodal tasks. However, beneath their seemingly impressive performance lies a critical limitation: MLLMs often fail to distinguish between task-relevant and task-irrelevant signals across modalities, leading to unreliable predictions [4, 5, 6]. A concrete manifestation of this limitation is their behavior in unimodal diagnostic settings, where only a single modality is required for correct reasoning. In such settings, MLLMs frequently underperform on pure visual recognition [7, 8] and textual reasoning [5, 9, 10], indicating that non-essential modalities can still influence the prediction process.

Recent studies have attributed this failure mode to a variety of learning challenges arising during the multimodal alignment process, including catastrophic forgetting of pretraining knowledge [7, 8, 9, 10], cross-modality knowledge conflict between visual and textual parametric memories [4, 5], and reliance on spurious cross-modal correlations rather than task-relevant grounding [11, 6, 12]. While each of these diagnoses captures a real aspect of the problem, they remain fragmented: each identifies a different cause located at a different stage of training or a different architectural component, and none specifies a common inference-time mechanism that governs how MLLMs identify and rely on task-relevant modalities during prediction. As a result, it remains unclear whether these failures

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

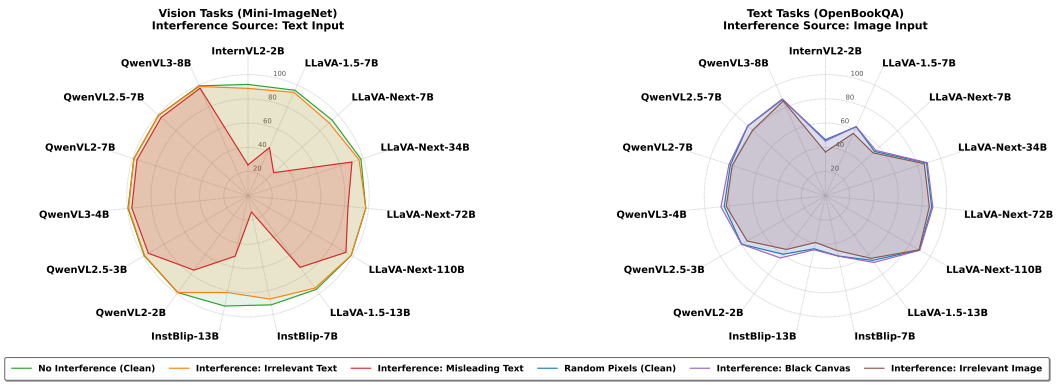


Figure 1: Modality Interference in MLLMs. We visualize the performance of 15 MLLMs using radar charts, where the polygon area signifies model capability. **Left (Vision Tasks):** When vision tasks are interfered with by *Misleading Text* (red), models exhibit a severe performance collapse, shrinking towards the center. **Right (Text Tasks):** Conversely, text reasoning tasks suffer from *Visual Noise* (brown/irrelevant image), causing a noticeable degradation compared to the no-interference baseline.

stem from a single underlying deficiency or from several independent problems that happen to co-occur. In this work, we provide evidence for the former and frame this underlying deficiency through the lens of *Cross-Modality Competency* [13], defined as the ability of an MLLM to fairly evaluate and integrate information across modalities. Current MLLMs lack the inference-time mechanisms to support this competency, and as a result, task-irrelevant modality signals improperly influence their predictions. We term this failure mode ***Modality Interference***, the concrete and measurable manifestation of insufficient cross-modality competency. Under this view, catastrophic forgetting [7, 8, 9, 10], knowledge conflict [4, 5], and spurious correlations [11, 6, 12] are no longer disjoint phenomena, but specific upstream causes that converge on the same observable symptom: predictions that shift when they should not.

To systematically diagnose modality interference, we design a perturbation-based evaluation grounded in the principles of causal intervention [14, 11]. We cast the multimodal prediction process as a causal graph (Figure 2) and define modality interference as the setting in which the task-irrelevant modality exerts a non-trivial causal effect on the prediction, measured by the prediction change induced by a controlled intervention on that modality. We instantiate this measurement across three regimes, each formulate as multiple-choice question answering: image-heavy tasks (e.g., image classification), text-heavy tasks (e.g., pure-text QA), and balanced multimodal tasks (e.g., VQA). Within each regime, we apply heuristic perturbations to the task-irrelevant modality, such as misleading textual descriptions for image-heavy tasks and distractor visual inputs for text-heavy ones, and measure how often the model's prediction flips relative to the clean input. Applying this protocol to a range of pretrained MLLMs reveals consistent failure patterns across model families and scales, as shown in Figure 1. Misleading textual descriptions trigger a sharp collapse on image-heavy tasks, while irrelevant visual content systematically disrupts text-heavy reasoning, indicating that current MLLMs are broadly vulnerable to modality interference.

These findings establish modality interference as a pervasive failure of cross-modality competency in current MLLMs, and motivate explicit mechanisms that target this failure during training. To this end, we propose a perturbation-aware fine-tuning framework that combines two complementary ingredients: a training mixture incorporating both heuristic noise and adversarial worst-case perturbations on the task-irrelevant modality, and a consistency regularizer (e.g., KL or JS divergence) that enforces output stability between clean and perturbed inputs. In summary, this paper makes three contributions. First, we identify the *Cross-Modality Competency Problem* as the deficiency underlying MLLMs' failures on cross-modal inputs, and introduce *Modality Interference* as its inference-time manifestation, unifying previously separate failure modes under one mechanism. Second, we propose a causal, perturbation-based evaluation protocol that quantifies modality interference across image-heavy, text-heavy, and balanced multimodal regimes, and use it to show that interference persists across model families and at scales up to 100B+ parameters. Third, we develop a dataset- and perturbation-agnostic fine-tuning framework that improves both unimodal robustness and multimodal accuracy across multiple MLLM families and scales, achieving Pareto-optimal gains.

2 Related Works

Improving Modality Alignment in Multimodal Language Models Recent studies have revealed that modality misalignment remains a key obstacle in MLLMs, leading to degraded performance on both image-heavy and text-heavy tasks. For visual understanding, catastrophic forgetting occurs when multimodal tuning overrides pretrained visual features [7, 8, 9]. In text-heavy scenarios, knowledge conflict [5] arises when inconsistent parametric knowledge from different modalities confuse reasoning. More recent analyses make this conflict more precise: [15] introduce controlled semantic interventions over image–text pairs and show that complementary inputs improve answers while contradictory ones sharply degrade accuracy and confidence; [16] construct MMA-Bench and find that current MLLMs systematically over-rely on text and collapse under conflicting cues; [17] further decompose conflict resolution into a case-specific reasoning-uncertainty gap and a stable inherent modality preference. Compositional benchmarks such as Winoground [18] and Vinoground [19] probe a complementary competency: both modalities are task-relevant and the model must bind their structure. mDPO [4] identifies language bias in training, where models fail to condition their responses on visual input. Some works attribute such issues to shallow fusion [9]—e.g., LLaVA [1] uses lightweight projectors to bridge vision and language spaces, leaving a representational gap and resulting in loosely coupled features [8, 5, 20]. Others highlight data limitations: even well-encoded visual features fail to support reasoning without adequate supervision to guide decoding [7]. Building on these diagnoses, recent works propose multiple solutions. MoF [8] mitigates this by fusing features from multiple vision encoders, while VLMClassifier [7] enhances recognition via vision-only finetuning, though it struggles with VQA due to lack of cross-modal alignment. CogVLM [9] introduces a visual expert module to improve vision-language integration. [16] additionally propose a modality alignment tuning strategy that teaches models when to prioritize, leverage, or ignore specific modality cues. VILA [10], QwenVL [3], and InternVL [21] incorporate text-only supervision in different ways to preserve or enhance language capabilities during multimodal training—through stage-wise separation, parallel preservation, and unified joint optimization, respectively. Similar patterns also arise in multimodal structural reasoning [22]. However, these approaches largely treat modality contributions implicitly, without explicitly modeling how multimodal inputs influence the model's decision process—even recent diagnostic frameworks [17, 15] stop at characterizing preference dynamics rather than translating them into training-time interventions.

Adversarial Robustness Across Modalities Adversarial perturbations threaten the reliability of both vision and text models by exploiting vulnerabilities in continuous image embeddings and discrete token spaces. In vision tasks, attacks like FGSM [23] and CW [24] first revealed the fragility of neural networks to imperceptible input changes. PGD [25] formalized this under a saddle-point framework, becoming the standard for adversarial training. AutoAttack [26] further unified strong attacks, including PGD variants, into a reliable benchmark. In text tasks, adversarial methods must contend with discrete inputs. TextFooler [27] substitutes key words with semantically similar ones to mislead predictions, while CodeAttack [28] adapts this idea to code-language models. More recently, PGD has been extended to LLMs via continuous relaxation [29], enabling efficient attacks in embedding space. PTP [30] applies PGD-style perturbations in the prompt embedding space to smooth training and enhance stability. Inspired by this, our work extends to the multimodal embedding space, enabling unified gradient-based control over both visual and textual inputs with unique masking strategy.

3 Causal Analysis on Modality Interference

Cross-Modality Competency Problems in Multimodal Large Language Models Competency problems describe scenarios where models rely on spurious correlations between isolated input features and output labels to make predictions, instead of leveraging meaningful interactions among multiple features [13]. We extend this concept to the multimodal setting by treating entire modalities (e.g., image X_I or text X_T) as structured feature sources. We define the *Cross-Modality Competency* as an ability for MLLM to fairly evaluate and integrate all modalities, identifying which ones carry task-relevant signals while ignoring misleading or irrelevant ones. In pure-text question answering with MLLMs, the model is provided with both a question and an image, although the image is not required to solve the task. When predictions are influenced by spurious visual cues correlated with certain answers, the model allows irrelevant modality signals to influence its reasoning and predictions, resulting in modality interference.

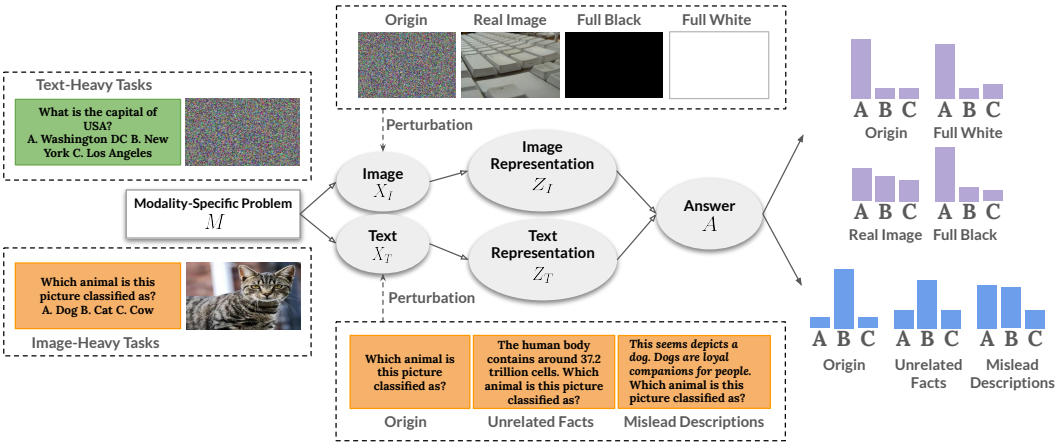


Figure 2: Causal graph illustrating modality interference in our perturbation-based evaluation analysis. Controlled interventions (heuristic) perturb either the image or text inputs, affecting their intermediate representations and ultimately the model prediction.

Perturbation-based Evaluation Experiment To systematically measure cross-modality competency, we propose a perturbation-based evaluation framework. The core idea is to inject controlled noise into the irrelevant modality and assess the model's robustness to such perturbations. Specifically, for image-heavy tasks, we perturb the text input by: (1)Prepending *unrelated scientific facts*; (2) Prepending *misleading descriptions* that falsely link incorrect options to the image content. For text-heavy tasks, we perturb the visual input by: (1) Attaching a *semantically meaningful real but irrelevant image*; (2) Substituting with a *full black* or *full white* canvas image. Models with strong modality selectivity should maintain high prediction consistency when irrelevant modality signals are perturbed. We select multiple image-heavy and text-heavy tasks for evaluation. Each task is framed as a multiple-choice classification problem, requiring the model to choose the correct option based on image and text modalities as input, with perturbations applied as described above. To validate robustness and generalization for our proposed mitigation method, we additionally consider real-world out-of-distribution perturbations reflecting practical noise encountered in deployed multimodal systems: (i) noisy OCR snippets for image-heavy tasks; and (ii) unrelated UI screenshots for text-heavy tasks. Details in Section 5, Section B.2, Table 7 and Table 9 and Table 10.

Causal Modeling of Modality Interference We formalize modality interference through a causal intervention perspective with a causal graph, as shown in Figure 2, where visual inputs (X_I) and textual inputs (X_T) are processed into their respective representations (Z_I, Z_T) before being fused to produce the final prediction (A). To study the model's reliance on different modalities, we introduce perturbations directly at the input level, serving as causal interventions [14] on X_I and X_T. Under ideal cross-modality competency, the model's prediction should primarily depend on the task-relevant pathway (e.g., $X_I \to Z_I \to A$ in image-heavy tasks, $X_T \to Z_T \to A$ in text-heavy tasks). Causal interventions at the input level allow us to diagnose whether the model improperly fuses irrelevant signals into its decision process. We use x'_I to denote an intervention on image X_I and use x'_T as an intervention on text X_T. Following Pearl's causal framework [14, 31], we quantify the impact of modality perturbations on model predictions by formalizing causal effects in our multimodal setting. Specifically, we define the pre-intervention prediction distribution as $P(A|X_I, X_T)$, and the post-intervention prediction distribution after applying a perturbation on X_I or X_T as $P'(A|\text{do}(X_I = x'_I)$ or $\text{do}(X_T = x'_T))$. The *do*-operation represents an intervention to specific modality, and the causal effect (CE) of an intervention is evaluated via a distance metric δ comparing P and P' as $\text{CE} = \delta(P, P')$. We assess the causal effect via prediction changes using $\delta_{\text{cp}}(P, P') := \mathbb{I}(a \neq a')$ in which $a = \arg\max_x P(x)$ is the predicted answer before intervention, $a' = \arg\max_x P'(x)$ is the predicted answer after intervention and $\mathbb{I}(\cdot)$ is the indicator function that outputs 1 if $a \neq a'$ and 0 otherwise. Thus, δ_{cp} captures whether the model's final decision A changes under perturbations to the input modality. In all interventions, a high value of δ_{cp} indicates the model's susceptibility to modality interference. We report δ_{cp} averaged over a dataset, i.e., the fraction of examples whose prediction changes (Section B.10). The causal effect δ_{cp} is the decision-level shadow of an underlying information-theoretic quantity, the conditional mutual information $I(Y; X_{\text{irr}} \mid X_{\text{rel}})$; the two perspectives are formally connected in Section C.

4 Methods

To mitigate modality interference and enhance cross-modality competency, we propose a unified perturbation-aware training framework that introduces interventions at both the input level (on X_I and X_T) and the representation level (on Z_I and Z_T) with consistency regularization. The framework can also be interpreted as a variational surrogate for a conditional Information Bottleneck objective (Section C). Overall pipeline is displayed in Figure 3.

4.1 Perturbation-based Data Augmentation

To increase causal robustness along the desired paths $X_I \to Z_I \to A$ (for image-heavy tasks) or $X_T \to Z_T \to A$ (for text-heavy tasks), we first adopt a causally grounded data augmentation method by augmenting each sample with both heuristic perturbations and training time adversarial perturbations.

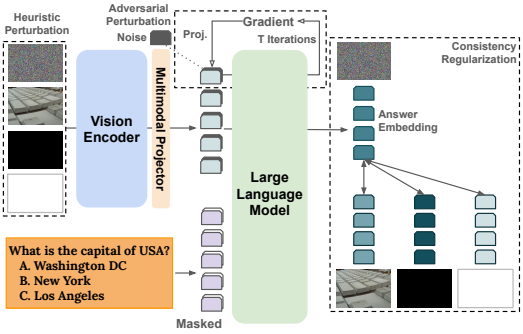


Figure 3: Overview of our proposed framework.

Mixture of Multi-Task Training data with Heuristic Perturbations Let $\mathcal{D} = (x_I, x_T, a)$ denote the full multimodal training dataset, where (x_I, x_T) are the image and text inputs and a is the ground-truth answer. The dataset $\mathcal{D}$ can be partitioned into three subsets based on tasks: 1) Image-heavy set $\mathcal{D}^{\text{img}}$: samples where visual input x_I is the dominant information source; 2) Text-heavy set $\mathcal{D}^{\text{text}}$: samples where textual input x_T provides the main reasoning signal; 3) VQA set $\mathcal{D}^{\text{vqa}}$: samples from vision-language datasets with naturally balanced multimodal dependencies. In practice, we transform these image-heavy and text-heavy datasets into VQA format to construct $\mathcal{D}^{\text{img}}$ and $\mathcal{D}^{\text{text}}$, and derive VQA samples from the supervised finetuning stage of each MLLM as $\mathcal{D}^{\text{vqa}}$. For each sample $(x_I, x_T, a) \in \mathcal{D}^{\text{img}} \cup \mathcal{D}^{\text{text}}$, we maintain its original version and apply heuristic perturbation to construct Origin Samples and Perturbation-Augmented Samples. Origin samples are used to reinforce the desired causal path (e.g., $X_I \to Z_I \to A$). Perturbation-augmented samples are variants of the same instance with perturbations applied to the irrelevant modality, which are denoted as $(x_I, \tilde{x}_T, a) \in \mathcal{D}^{\text{img}}_{\text{pert}}$ and $(\tilde{x}_I, x_T, a) \in \mathcal{D}^{\text{text}}_{\text{pert}}$ where $\tilde{x}_T$ and $\tilde{x}_I$ are perturbed text and image respectively. Together, the augmented dataset can be written as:

$$\mathcal{D}^{\text{AUG}} = \mathcal{D}^{\text{img}} \cup \mathcal{D}^{\text{img}}_{\text{pert}} \cup \mathcal{D}^{\text{text}} \cup \mathcal{D}^{\text{text}}_{\text{pert}} \cup \mathcal{D}^{\text{VQA}}. \tag{1}$$

We sample N_{img} and N_{text} examples from $\mathcal{D}^{\text{img}}$ and $\mathcal{D}^{\text{text}}$ to construct $\mathcal{B}^{\text{img}}_{\text{orig}}$ and $\mathcal{B}^{\text{text}}_{\text{orig}}$ respectively, and the remaining N_{vqa} examples are VQA samples from $\mathcal{D}^{\text{vqa}}$ to construct $\mathcal{B}^{\text{vqa}}$. With dynamically generated perturbed variants for each sample, the final training batch is:

$$\mathcal{B} = \mathcal{B}^{\text{img}}_{\text{orig}} \cup \mathcal{B}^{\text{img}}_{\text{pert}} \cup \mathcal{B}^{\text{text}}_{\text{orig}} \cup \mathcal{B}^{\text{text}}_{\text{pert}} \cup \mathcal{B}^{\text{vqa}}, \tag{2}$$

where $\mathcal{B}^{\text{img}}_{\text{pert}}$ and $\mathcal{B}^{\text{text}}_{\text{pert}}$ are perturbation-augmented variants generated from $\mathcal{B}^{\text{img}}_{\text{orig}}$ and $\mathcal{B}^{\text{text}}_{\text{orig}}$, respectively. For the full training batch $\mathcal{B}$, which includes both original and perturbed samples, we define the supervised loss $\mathcal{L}_{\text{sft}}$ as the cross-entropy loss computed over all answer tokens in the ground-truth sequences. Let $\mathcal{L}_{\text{cls}}(x_I, x_T, a)$ denote the standard autoregressive loss for a sample (x_I, x_T, a), then:

$$\mathcal{L}_{\text{sft}} = \frac{1}{|\mathcal{B}|} \sum_{(x_I, x_T, a) \in \mathcal{B}} \mathcal{L}_{\text{cls}}(x_I, x_T, a). \tag{3}$$

Adversarial Perturbation with Cross-Modality Masking While heuristic perturbations simulate realistic but limited modality noise at the input level, they may not fully capture the worst-case failure modes of MLLMs, especially under complex spurious alignments in the representation space. To overcome this limitation, we introduce a stronger and more generalizable intervention through adversarial training. These perturbations simulate worst-case alignment disruptions during training, serving as targeted interventions on latent nodes (Z_I, Z_T) to reveal the *Direct Causal Effect* (DCE) of irrelevant modalities on A. By optimizing the model under such adversarial conditions, we reduce the model's reliance on spurious cross-modal signals and reinforce task-relevant causal pathways.

Inspired by PGD [25, 32], we design a tailored perturbation strategy for multimodal token embeddings (Z_I, Z_T). Unlike standard PGD that applies coarse sign-based updates, our method introduces two modifications: (1) *Modality-specific perturbation masking*, which restricts perturbations to task-irrelevant modalities via a binary mask, thereby transforming noise into targeted causal probes rather than indiscriminate corruption. (2) *Raw-gradient updates*, where we remove the sign operator and apply the raw gradient directly. Unlike the sign operator, which pushes every coordinate to the ℓ_∞ boundary under our step size ($\alpha \gg \epsilon$), raw gradients preserve each coordinate's relative sensitivity, while the projection in Eq. (5) bounds the perturbation magnitude (Section B.8). Formally, we construct perturbations $\delta = (\delta_I, \delta_T)$ in the latent space that maximize the model's predictive loss:

$$\delta = \underset{\|\delta\|_\infty \leq \epsilon}{\operatorname{argmax}} \mathcal{L}_{\text{cls}}(f(Z_I + \delta_I, Z_T + \delta_T)), \tag{4}$$

where ϵ bounds the perturbation strength and f is the prediction function. We optimize δ through n raw-gradient steps, updating at each step t as:

$$\delta^{(t+1)} = \Pi_{\|\delta\|_\infty \leq \epsilon}\Big(\delta^{(t)} + \alpha \cdot \nabla_\delta \mathcal{L}_{\text{cls}}(f(Z + \delta^{(t)}))\Big), \tag{5}$$

where α is the step size, Π projects the noise into the ℓ_∞ ball, and $Z = [Z_I; Z_T]$ is the concatenated embedding. In practice, we integrate a modality-specific binary mask $M \in \{0,1\}^{L\times d}$, where L is the sequence length and d the hidden dimension, ensuring that perturbations only affect task-irrelevant tokens. Given multimodal embeddings $\mathbf{E} \in \mathbb{R}^{L\times d}$, the perturbed embeddings are:

$$\tilde{\mathbf{E}} = \mathbf{E} + \delta \odot M, \tag{6}$$

with $\odot$ denoting element-wise masking. We initialize δ with Gaussian noise $\mathcal{N}(0, \epsilon^2)$ and update it for T steps. The final adversarial objective is $\mathcal{L}_{\text{adv}} = \mathcal{L}_{\text{cls}}(f(\tilde{\mathbf{E}}))$.

4.2 Consistency Regularization under Perturbations

While perturbation-based augmentation exposes the model to diverse interventions, it does not constrain how intermediate representations (Z_I, Z_T) should respond, and even small changes in the task-irrelevant modality may cause undesirable shifts in fused features. To address this, we introduce a consistency regularization strategy that enforces *output stability* between original and perturbed inputs, serving as an indirect constraint on Z_I and Z_T to mitigate modality interference. By minimizing the divergence between the prediction distributions of original and perturbed inputs, the model is encouraged to maintain invariant behavior along the task-relevant causal paths. Formally, given an original input x and its perturbed counterpart $\tilde{x}$, with predictive distributions $p_\theta(A|x)$ and $p_\theta(A|\tilde{x})$, the consistency loss follows the general form:

$$\mathcal{L}_{\text{consistency}} = \text{Consistency}\big(p_\theta(A|x) \,\|\, p_\theta(A|\tilde{x})\big). \tag{7}$$

In practice, we instantiate this by applying distributional divergence (e.g., KL or JS) at the token level. Let $l^{\text{orig}}, l^{\text{pert}} \in \mathbb{R}^{L_A\times V}$ denote the pre-softmax logits for the original and perturbed samples, where L_A is the number of answer tokens and V the vocabulary size. Using KL divergence with temperature τ as an example, the loss becomes (equally apply to image-heavy and text-heavy tasks):

$$\mathcal{L}_{\text{consistency}} = \frac{1}{L_A}\sum_{i=1}^{L_A}\sum_{v=1}^{V} \text{softmax}\Big(\tfrac{l_i^{\text{orig}}}{\tau}\Big)_v \cdot \log \frac{\text{softmax}\Big(\tfrac{l_i^{\text{orig}}}{\tau}\Big)_v}{\text{softmax}\Big(\tfrac{l_i^{\text{pert}}}{\tau}\Big)_v}. \tag{8}$$

4.3 Final Training Objective

Our final training objective integrates both perturbation-based data augmentation and consistency regularization into a unified framework. For each batch, we begin with a set of original samples $\mathcal{B}_{\text{orig}}$ and dynamically construct their heuristic perturbed counterparts $\mathcal{B}_{\text{pert}}$ via input-level augmentations. We then apply adversarial perturbations on both $\mathcal{B}_{\text{orig}}$ and $\mathcal{B}_{\text{pert}}$, and enforce consistency between the original and all perturbed samples. The overall loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sft}} + \mathcal{L}_{\text{adv}} + \lambda_{\text{cons}} \cdot \mathcal{L}_{\text{consistency}}, \tag{9}$$

where $\mathcal{L}_{\text{sft}}$ is the supervised loss computed over $\mathcal{B}$, $\mathcal{L}_{\text{adv}}$ is the adversarial loss computed on all adversarial perturbed samples and $\mathcal{L}_{\text{consistency}}$ is the consistency loss between original and all perturbed sample pairs. By aligning all three losses with the causal structure of multimodal reasoning, we systematically mitigate modality interference and improve cross-modality competency in MLLMs.

Table 1: Evaluation of unimodal and multimodal tasks across multiple methods. For unimodal datasets, we report accuracy on origin setting (Orig) and the worst-performing perturbation (Perturbed). For VQA datasets, we report accuracy on the original setting. The best accuracy is marked in bold.

Model Settings	**Mini-ImageNet**		**Caltech-101**		**OpenBookQA**		**MMLU**		**ScienceQA**	**MM-Bench**	**Seed-Bench**
	Orig	Perturbed	Orig	Perturbed	Orig	Perturbed	Orig	Perturbed	Accuracy	Accuracy	Accuracy
LLaVA-1.5-7B	95.3	43.5	97.0	57.4	62.4	56.4	46.3	45.2	64.5	64.3	63.4
+ CoT [33]	81.7	28.9	80.9	36.0	38.8	38.9	39.6	38.7	64.7	65.2	64.1
+ VLMClassifier [7]	15.6	0.0	15.1	0.0	61.8	61.2	47.4	47.8	61.8	35.8	36.0
+ I-MoF [8]	91.8	41.2	97.0	58.1	67.1	63.8	46.5	40.0	62.1	65.2	64.2
+ CausalMM [12]	95.0	47.2	97.3	64.5	61.8	56.6	47.0	47.0	63.9	64.3	63.7
+ VCD [34]	94.3	50.9	96.8	63.3	61.7	56.4	46.9	46.6	63.0	63.8	63.8
Ours	**98.6**	**98.4**	**99.3**	**98.9**	**81.8**	**81.0**	**51.5**	**51.0**	**67.8**	**73.1**	**64.6**
LLaVA-1.5-13B	95.6	73.0	97.9	77.4	65.9	63.8	51.8	50.8	65.8	72.1	64.5
+ CoT [33]	92.9	62.8	96.6	67.8	55.6	53.0	47.3	45.5	65.6	70.2	64.9
+ VLMClassifier [7]	25.8	0.0	24.9	0.0	66.8	65.3	52.9	51.3	66.7	56.2	55.3
+ I-MoF [8]	93.9	70.1	97.8	80.9	69.2	64.5	46.2	42.1	**66.8**	73.0	66.6
+ CausalMM [12]	95.3	75.1	97.0	83.5	67.5	58.1	51.9	50.3	64.1	72.0	64.1
+ VCD [34]	95.2	75.7	97.8	82.4	65.7	63.4	51.9	50.6	66.3	70.9	63.3
Ours	**98.5**	**98.4**	**99.2**	**98.6**	**83.0**	**82.1**	**56.6**	**55.8**	65.3	**73.7**	**69.7**

5 Experiments

Models. We conduct experiments on three MLLM families with different parameter size: Qwen2.5-vl-3b [3], LLaVA-1.5-7B & LLaVA-1.5-13B [35] and InstructBLIP-Vicuna-7B [2]. Following LLaVA and Qwen-VL, we freeze the vision encoder and train the multimodal projector and language model; for InstructBLIP, we instead freeze both the vision encoder and Q-Former, fine-tuning only the projection layer and language model (see Section B.1).

Baselines. We compare against four families of mitigation strategies for modality interference, ordered by the extent of intervention. **(i) Prompting.** *Chain-of-Thought (CoT)* [33] elicits step-by-step reasoning through prompt design alone, without modifying weights (prompt in Section B.4). **(ii) Contrastive decoding.** These methods subtract modality-prior bias at decoding time by contrasting against a counterfactual prediction, with no weight changes. *CausalMM* [12] edits visual self-attention as a counterfactual intervention; *VCD* [34] perturbs the input image with diffusion-style Gaussian noise. Both combine the original and counterfactual logits via $\ell_{\text{cm}} = (1+\gamma)\ell_{\text{orig}} - \gamma\ell_{\text{cf}}$. **(iii) Architectural modification.** *I-MoF* [8] augments LLaVA-1.5-13B by interleaving CLIP [36] and DINOv2 [37] visual tokens to combine contrastive and self-supervised features. *VLMClassifier* [7] fine-tunes the model on ImageNet [38] combined with LLaVA-Instruct [1] for image classification.

Datasets. We evaluate models on benchmarks covering three task types: *(i) Image-heavy tasks*: Mini-ImageNet [39] and Caltech-101 [40], used in both training and evaluation; *(ii) Text-heavy tasks*: OpenBookQA [41] and MMLU [42], used in both training and evaluation; *(iii) VQA tasks*: For training, we use LLaVA-Instruct-dataset [1] as the instruction-tuning dataset for related models. For InstructBLIP, we additionally use TextCaps [43] following its original training configuration. For Qwen2.5-VL, whose instruction-tuning data is proprietary, we adopt LLaVA-Instruct as a standardized alternative. For evaluation, we adopt three multiple-choice VQA benchmarks: ScienceQA-IMG [44], MM-Bench-EN [45], and Seed-Bench-IMG [46]. For ScienceQA and Seed-Bench, we only include examples with image context. For MM-Bench, we use the English version. All datasets are converted into a unified multiple-choice VQA format. We report the accuracy of model predictions and quantify the causal effect with the prediction change rate δ_{cp}. All models are fine-tuned for 1 epoch with a fixed batch size N_{batch}. All the hyperparameters are listed in Section D. For each dataset, results are averaged over multiple independent runs. Following standard practice, inference is performed with deterministic decoding (temperature fixed at 0).

Achieving Pareto-Optimality Across Unimodal and Multimodal Tasks. Table 1 shows that our method outperforms every baseline on both LLaVA-1.5-7B and LLaVA-1.5-13B. The pattern of failures is informative: each prior method implicitly diagnoses a different root cause of modality interference, and each diagnosis misses the target. CoT presumes the model fails because it does not reason long enough, but longer chains give the spurious modality more room to be rationalized, and perturbed accuracy drops below vanilla. Decoding-time methods (CausalMM, VCD) subtract a counterfactual logit term to remove modality-prior bias without parameter updates, and therefore cannot reshape the cross-modal representations that produce interference in the first place. Architectural fixes target representations but not modality reliance: I-MoF strengthens the visual backbone yet still degrades on perturbed image tasks and even harms text-only reasoning, while VLMClassifier specializes for image

Table 2: Multimodal reasoning accuracy (%) on VQA benchmarks under various ablation study configurations. The best accuracy is marked in bold. Arrows (↑ / ↓) indicate relative changes compared to the Vanilla baseline of each model. Overall performance is computed as a weighted average across datasets, with weights proportional to each dataset's test size. (full results in Table 6.)

Model	Method	ScienceQA-IMG	MM-Bench-EN	Seed-Bench-IMG	VQA Overall
Qwen2.5-VL family [3]					
Qwen2.5-VL-3B [3]	Vanilla	63.0 (100%)	81.8 (100%)	72.3 (100%)	72.5 (100%)
	FFT with D^{VQA}	72.2 (↑14.6%)	82.0 (↑0.2%)	74.7 (↑3.3%)	75.5 (↑3.9%)
	FFT with D^{AUG}	73.6 (↑16.8%)	81.7 (↓0.1%)	75.3 (↑4.1%)	76.2 (↑5.1%)
	+ KL	72.8 (↑15.6%)	80.9 (↓1.1%)	74.9 (↑3.6%)	75.9 (↑4.7%)
	+ ADV	73.1 (↑16.0%)	81.7 (↓0.1%)	75.3 (↑4.1%)	76.1 (↑5.0%)
	Ours	73.8 (↑17.1%)	82.1 (↑0.1%)	75.5 (↑4.4%)	**76.4 (↑5.4%)**
Qwen2.5-VL-7B	Vanilla	85.1 (100%)	86.6 (100%)	77.0 (100%)	80.3 (100%)
	FFT with D^{VQA}	81.4 (↓4.3%)	86.4 (↓0.2%)	76.8 (↓0.3%)	79.5 (↓1.0%)
	FFT with D^{AUG}	83.2 (↓2.2%)	86.1 (↓0.6%)	77.0 (0.0%)	79.9 (↓0.5%)
	+ JS	85.5 (↑0.5%)	85.9 (↓0.8%)	77.0 (0.0%)	80.3 (0.0%)
	+ ADV	85.2 (↑0.1%)	85.8 (↓0.9%)	76.6 (↓0.5%)	79.9 (↓0.5%)
	Ours	83.9 (↓1.4%)	86.4 (↓0.2%)	78.3 (↑1.7%)	**80.9 (↑0.7%)**
LLaVA-1.5 family [35]					
LLaVA-1.5-7B [35]	Vanilla	64.5 (100%)	64.3 (100%)	63.4 (100%)	63.8 (100%)
	FFT with D^{VQA}	61.6 (↓4.5%)	71.4 (↑11.0%)	62.4 (↓1.6%)	64.0 (↑0.3%)
	FFT with D^{AUG}	65.4 (↑1.4%)	71.1 (↑10.6%)	63.9 (↑0.8%)	65.6 (↑2.8%)
	+ JS	63.8 (↓1.1%)	73.5 (↑14.3%)	63.7 (↑0.5%)	65.6 (↑2.8%)
	+ ADV	66.7 (↑3.4%)	71.4 (↑11.0%)	63.6 (↑0.3%)	65.7 (↑3.0%)
	Ours	67.8 (↑5.1%)	73.1 (↑13.7%)	64.6 (↑1.9%)	**66.8 (↑4.7%)**
LLaVA-1.5-13B	Vanilla	65.8 (100%)	72.1 (100%)	64.5 (100%)	66.2 (100%)
	FFT with D^{VQA}	60.8 (↓7.6%)	73.6 (↑2.1%)	64.9 (↑0.6%)	65.8 (↓0.6%)
	FFT with D^{AUG}	63.5 (↓3.5%)	75.0 (↑4.0%)	65.7 (↑1.9%)	67.1 (↑1.4%)
	+ KL	62.5 (↓5.0%)	74.6 (↑3.5%)	67.6 (↑4.8%)	68.0 (↑2.7%)
	+ ADV	57.6 (↓12.5%)	74.2 (↑2.9%)	68.3 (↑5.9%)	67.5 (↑2.0%)
	Ours	65.3 (↓0.8%)	73.6 (↑2.1%)	69.7 (↑8.1%)	**68.4 (↑3.3%)**
InstructBLIP family [2]					
InstructBLIP-7B	Vanilla	52.1 (100%)	65.5 (100%)	54.8 (100%)	56.4 (100%)
	FFT with D^{VQA}	61.8 (↑18.6%)	68.4 (↑4.4%)	57.6 (↑5.1%)	60.4 (↑7.1%)
	FFT with D^{AUG}	65.4 (↑25.5%)	71.1 (↑8.5%)	61.2 (↑11.7%)	63.9 (↑13.3%)
	+ KL	65.9 (↑26.5%)	71.7 (↑9.5%)	60.9 (↑11.1%)	63.9 (↑13.3%)
	+ ADV	66.0 (↑26.7%)	69.0 (↑5.3%)	61.8 (↑12.8%)	64.0 (↑13.5%)
	Ours	64.0 (↑22.8%)	71.4 (↑9.0%)	63.0 (↑15.0%)	**64.8 (↑14.9%)**
InstructBLIP-13B [2]	Vanilla	65.8 (100%)	72.1 (100%)	63.8 (100%)	65.8 (100%)
	FFT with D^{VQA}	61.7 (↓6.2%)	68.5 (↓5.0%)	56.0 (↓12.2%)	59.4 (↓9.7%)
	FFT with D^{AUG}	66.6 (↑1.2%)	71.5 (↓0.8%)	64.0 (↑0.3%)	65.9 (↑0.2%)
	+ KL	67.2 (↑2.2%)	72.6 (↑0.7%)	64.0 (↑0.3%)	66.2 (↑0.6%)
	+ ADV	66.1 (↑0.5%)	74.0 (↑2.6%)	64.1 (↑0.5%)	66.4 (↑0.9%)
	Ours	68.2 (↑3.6%)	74.4 (↑3.2%)	65.5 (↑1.0%)	**67.7 (↑2.8%)**

classification through coarse task-level gating, which collapses both unperturbed image accuracy and multimodal accuracy. In contrast, our method addresses interference at its source: adversarial perturbations on the task-irrelevant modality, combined with consistency regularization, supply the training signal these baselines lack and directly reshape the model's modality reliance. The result is uniform improvement: across both backbones, our method improves all unimodal benchmarks under both clean and perturbed inputs and all three VQA benchmarks, achieving *Pareto-optimality* gains.

Beyond accuracy, our framework directly reduces the causal effect defined in Section 3: δ_{cp} decreases under every evaluated intervention across all three backbones, e.g.,from 9.0% to 4.2% for QwenVL3-4B under unrelated real images on OpenBookQA (Table 13).

5.1 Ablation Studies

To evaluate the effectiveness of each component in our framework, we conduct a comprehensive ablation study across multiple models and scales. We compare the pretrained models with the following strategies: *FFT with* D^{VQA} (standard finetuning on VQA data), *FFT with* D^{AUG} (supervised

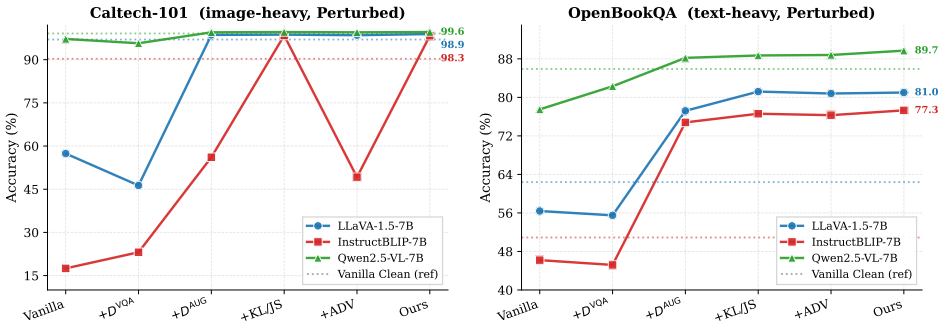


Figure 4: Ablation study on Caltech-101 and OpenBookQA Perturbed accuracy across the six-step ablation. Dotted horizontal lines mark each model's Vanilla *Clean* accuracy. Full results in Table 5.

Table 3: Out-of-distribution (OOD) robustness under unseen perturbation types and unseen datasets. **Seen datasets:** Mini-ImageNet, Caltech-101, OpenBookQA, MMLU (perturbations OCR/Screenshot are unseen). **Unseen datasets:** Food-101, ARC Challenge (neither used during training).

Model	Method	Image-heavy						Text-heavy					
		Mini-ImageNet		Caltech-101		Food-101		OpenBookQA		MMLU		ARC Challenge	
		Orig	OCR	Orig	OCR	Orig	OCR	Orig	Scrn.	Orig	Scrn.	Orig	Scrn.
LLaVA-1.5-7B	Vanilla	95.3	88.9	97.0	92.8	90.2	83.6	62.4	50.2	46.3	44.8	53.4	51.8
	$+D^{\text{AUG}}$	98.2	98.0	98.5	98.0	91.3	83.5	78.6	78.0	51.1	50.6	55.7	53.8
	+ADV	98.7	98.2	98.7	98.4	91.8	86.7	81.7	81.3	50.6	51.3	57.9	57.4
	Ours	**98.6**	**98.2**	**99.3**	**99.0**	**92.1**	**89.5**	**81.8**	**81.2**	**51.5**	**51.3**	**62.8**	**62.6**
InstructBLIP-7B	Vanilla	92.0	81.4	90.3	83.6	67.9	48.0	50.9	40.0	35.3	34.9	33.7	36.5
	$+D^{\text{AUG}}$	98.5	95.7	99.0	98.4	55.5	56.5	75.0	74.4	50.0	49.2	27.3	30.7
	+ADV	98.7	97.5	99.5	98.8	93.2	85.9	76.8	76.2	49.3	49.3	50.1	44.5
	Ours	**98.4**	**97.3**	**99.2**	**99.0**	**95.6**	**88.1**	**79.0**	**77.8**	**50.2**	**49.2**	**57.6**	**57.0**
Qwen2.5-VL-7B	Vanilla	99.3	99.2	99.1	99.2	97.1	96.7	85.9	69.3	69.3	57.0	**85.6**	79.6
	$+D^{\text{AUG}}$	99.6	99.5	**99.7**	99.3	97.1	96.5	90.2	85.8	70.4	63.7	85.0	80.2
	+ADV	99.4	99.5	99.6	**99.6**	97.3	96.9	91.7	**91.5**	**70.4**	69.7	85.0	81.2
	Ours	**99.6**	**99.5**	99.6	99.5	**97.5**	**97.4**	**90.9**	90.7	69.8	**69.7**	84.1	**82.6**

finetuning on mixed multi-task datasets with heuristic perturbations), *FFT+KL/JS* (adding consistency regularization on KL or JS divergence), *FFT+ADV* (FFT with heuristic & adversarial perturbations), and *Ours* (combining both perturbation-based data augmentation and consistency regularization). All configurations are trained under matched compute (same batch size and step count); see Section B.7 for the full statement. Table 2 presents overall VQA performance, and Figure 4 shows model robustness under unimodal settings. (Table 4 reports results with all perturbations.) Together, these results show the effectiveness of our method in improving both general VQA accuracy and robustness under modality interference. Across all model families (Qwen, InstructBLIP, LLaVA) and scales (3B/7B/13B), our method achieves the best overall performance, improving accuracy on both unimodal and multimodal benchmarks. The gains are particularly pronounced on perturbed inputs (e.g., $17.5\% \rightarrow 99.0\%$ with InstructBLIP-7B on Caltech-101), confirming that the framework directly mitigates modality interference rather than only improving general accuracy. We also extend evaluation from MCQA to free-form QA and additional diagnostic experiments, including a no-mask ablation, which exhibit consistent robustness trends (see Sections B.5 and B.9).

We observe consistent improvements across both unimodal and multimodal tasks when moving from *FFT w/* D^{VQA} to *FFT w/* D^{AUG}, highlighting the importance of incorporating modality-specific supervision and heuristic perturbations. Building upon this, adding consistency regularization yields further gains by stabilizing model predictions under controlled perturbations on X_I or X_T. Both KL and JS objectives lead to similar improvements, suggesting that the model equally benefits from all heuristic perturbations regardless of anchor choice. To further validate generalization, we introduce two real-world out-of-distribution perturbations at test time: (i) noisy OCR snippets sampled from FUNSD [47] as irrelevant text into image-heavy tasks; and (ii) unrelated UI screenshots from RICO [48] as distractor images in text-heavy tasks. Beyond existing datasets, we further evaluate on Food-101 [49] (unseen image-heavy dataset) and ARC-Challenge [50] (unseen text-heavy dataset), neither of which is used during training. As shown in Table 3, our method significantly improves robustness under these unseen perturbations with consistent gains (e.g., on InstructBLIP-7B, 83.6%

$\rightarrow$ 99.0% under OCR noise on Caltech-101). Crucially, results on new datasets confirm our method's strong OOD generalization. Under misleading text (seen perturbation) and OCR-style noise (unseen perturbation), vanilla models suffer severe degradation, while our method preserves accuracy close to clean-baseline performance. In contrast, heuristic data augmentation alone fails to generalize to unseen noise, and may even degrade performance (e.g., on InstructBLIP-7B, heuristic augmentation degrades performance on Food-101 under clean inputs (67.9% vs. 55.5%). These results demonstrate that **the modest overhead of adversarial training** (see Section E) **yields substantial gains in out-of-domain generalization** (full results in Table 9 and Table 10).

6 Conclusion

We identify and formalize *Modality Interference* as a concrete, measurable manifestation of insufficient *Cross-Modality Competency* in MLLMs, defined as the ability to fairly evaluate and integrate information across modalities. Through a perturbation-based evaluation grounded in causal intervention, we show that current MLLMs, including state-of-the-art ones, systematically fail under task-irrelevant but misleading modality signals, exposing a vulnerability that is not resolved by scale alone. Building on this diagnosis, we propose a perturbation-aware fine-tuning framework that combines heuristic and adversarial perturbations on the task-irrelevant modality with output-level consistency regularization between clean and perturbed inputs. The framework directly reshapes how the model relies on each modality at inference time, and yields consistent improvements in both unimodal robustness and multimodal accuracy across diverse architectures, scales, and task regimes, achieving Pareto-optimal trade-offs over prior baselines. More broadly, our results suggest that strengthening cross-modality competency requires explicit training-time signals that target modality reliance, beyond what better alignment objectives or representation-level fixes can provide.

Acknowledgments

We appreciate the reviewers for their constructive comments and suggestions. Rui Cai and Zhe Zhao were supported by NSF grant OAC 2531126 and a Google Research Award. Xiaofei Wen and Muhao Chen were supported by a Noyce Initiative Grant, the NSF of the United States under Grants ITE 2333736 and OAC 2531126, the NIH 1R01CA323939-01, and equipment gifts from NVIDIA and DELL.

References

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

[2] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500, 2023.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923, 2025.

[4] Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mdpo: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8078–8088, 2024.

[5] Tinghui Zhu, Qin Liu, Fei Wang, Zhengzhong Tu, and Muhao Chen. Unraveling cross-modality knowledge conflicts in large vision-language models. *ArXiv*, abs/2410.03659, 2024.

[6] Parsa Hosseini, Sumit Nawathe, Mazda Moayeri, Sriram Balasubramanian, and Soheil Feizi. Seeing what's not there: Spurious correlation in multimodal llms. *ArXiv*, abs/2503.08884, 2025.

[7] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? *ArXiv*, abs/2405.18415, 2024.

[8] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9568–9578, 2024.

[9] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. *ArXiv*, abs/2311.03079, 2023.

[10] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pretraining for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699, 2024.

[11] Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective. *ArXiv*, abs/2403.18346, 2024.

[12] Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[13] Matt Gardner, William Merrill, Jesse Dodge, Matthew E. Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. Competency problems: On finding and removing artifacts in language data. *ArXiv*, abs/2104.08646, 2021.

[14] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82:669–688, 1995.

[15] Kenza Amara, Lukas Klein, Carsten T. Lüth, Paul F. Jäger, Hendrik Strobelt, and Mennatallah El-Assady. Why context matters in vqa and reasoning: Semantic interventions for vlm input modalities. *ArXiv*, abs/2410.01690, 2024.

[16] Tianle Chen, Chaitanya Chakka, Arjun Akula, Xavier Thomas, and Deepti Ghadiyaram. Some modalities are more equal than others: Decoding and architecting multimodal integration in mllms. *ArXiv*, abs/2511.22826, 2025.

[17] Zhuoran Zhang, Tengyue Wang, Xilin Gong, Yang Shi, Haotian Wang, Di Wang, and Lijie Hu. When modalities conflict: How unimodal reasoning uncertainty governs preference dynamics in mllms. *ArXiv*, abs/2511.02243, 2025.

[18] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5228–5238. IEEE, 2022.

[19] Jianrui Zhang, Mu Cai, and Yong Jae Lee. Vinoground: Scrutinizing lmms over dense temporal reasoning with short videos. *arXiv preprint arXiv:2410.02763*, 2024.

[20] Yaqi Zhao, Yuanyang Yin, Lin Li, Mingan Lin, Victor Shea-Jay Huang, Siwei Chen, Weipeng Chen, Baoqun Yin, Zenan Zhou, and Wentao Zhang. Beyond sight: Towards cognitive alignment in lvlm via enriched visual knowledge. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24950–24959, 2025.

[21] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[22] Rui Cai, Shichao Pei, and Xiangliang Zhang. Zero-shot relational learning for multimodal knowledge graphs. In *2024 IEEE International Conference on Big Data (BigData)*, pages 499–508. IEEE, 2024.

[23] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[24] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57. IEEE Computer Society, 2017.

[25] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[26] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 2206–2216. PMLR, 2020.

[27] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8018–8025. AAAI Press, 2020.

[28] Akshita Jha and Chandan K. Reddy. Codeattack: Code-based adversarial attacks for pre-trained programming language models. In Brian Williams, Yiling Chen, and Jennifer Neville, editors, *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 14892–14900. AAAI Press, 2023.

[29] Simon Geisler, Tom Wollschläger, M. H. I. Abdalla, Johannes Gasteiger, and Stephan Günnemann. Attacking large language models with projected gradient descent. *CoRR*, abs/2402.09154, 2024.

[30] Lichang Chen, Jiuhai Chen, Heng Huang, and Minhao Cheng. PTP: boosting stability and performance of prompt tuning with perturbation-based regularizer. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 13512–13525. Association for Computational Linguistics, 2023.

[31] Probabilistic and causal inference: The works of judea pearl. *Probabilistic and Causal Inference*, 2022.

[32] Lichang Chen, Heng Huang, and Minhao Cheng. Ptp: Boosting stability and performance of prompt tuning with perturbation-based regularizer. *ArXiv*, abs/2305.02423, 2023.

[33] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[34] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Li Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13872–13882, 2023.

[35] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26286–26296, 2023.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[37] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[38] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[40] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178, 2004.

[41] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Conference on Empirical Methods in Natural Language Processing*, 2018.

[42] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Xiaodong Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *ArXiv*, abs/2009.03300, 2020.

[43] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020.

[44] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and A. Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *ArXiv*, abs/2209.09513, 2022.

[45] Yuanzhan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, 2023.

[46] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *ArXiv*, abs/2307.16125, 2023.

[47] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE, 2019.

[48] Biplab Deka, Zifeng Huang, Chad Franzen, Joshua Hibschman, Daniel Afergan, Yang Li, Jeffrey Nichols, and Ranjitha Kumar. Rico: A mobile app dataset for building data-driven design applications. In *Proceedings of the 30th annual ACM symposium on user interface software and technology*, pages 845–854, 2017.

[49] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014.

[50] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[51] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[52] Haotian Liu, Chunyuan Li, Yuheng Li, and Jianfeng Gao. Llava-next: Stronger llms with better visual instruction tuning, 2024.

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

[54] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xiong-Hui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Rongyao Fang, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Qidong Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Li Ying Meng, Xuancheng Ren, Xin yi Ren, Sibo Song, Yu-Chen Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yihe Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Botao Zheng, Humen Zhong, Jingren Zhou, Fanxi Zhou, Jingren Zhou, Yuanzhi Zhu, and Keming Zhu. Qwen3-vl technical report. *ArXiv*, abs/2511.21631, 2025.

[55] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

[56] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *ArXiv*, physics/0004057, 2000.

[57] DeepSpeed Contributors. Deepspeed. `https://github.com/deepspeedai/DeepSpeed`, 2024. Accessed: 2025-05-14.

A Appendix Summary

This appendix provides supplementary results, analyses, and implementation details supporting our study of modality interference in MLLMs. It is organized as follows.

Full Ablation Results. Table 4 reports unimodal accuracy for every ablation configuration under each individual perturbation (UF/MD for image-heavy tasks; RP/RI/FB/FW for text-heavy tasks) across Qwen2.5-VL-3B/7B, LLaVA-1.5-7B/13B, and InstructBLIP-7B/13B. Table 5 summarizes the same configurations by the original and worst-case perturbed accuracy together with overall VQA accuracy, and Table 6 gives the per-benchmark VQA results (ScienceQA-IMG, MM-Bench-EN, Seed-Bench-IMG).

Detailed Experimental Results (Section B):

- *Finetuning strategies* (Section B.1): our choice to freeze the Q-Former in InstructBLIP-based models.
- *Perturbation-based evaluation of vanilla MLLMs* (Section B.2): full results for 15 pretrained MLLMs from 2B to 110B parameters (Table 7, Figure 5), including the scaling trend within the LLaVA-Next family.
- *KL vs. JS consistency regularization*: a comparison of the two divergence choices.
- *Chain-of-Thought prompting* (Section B.4): the CoT prompt used for the CoT baseline in the main paper.
- *Free-form VQA* (Section B.5): results on the generative TextVQA benchmark (Table 8).
- *Out-of-distribution generalization*: robustness to unseen perturbation types, namely OCR noise from FUNSD [47] and UI screenshots from RICO [48] (Table 9), and to unseen datasets, namely Food-101 [49] and ARC-Challenge [50] (Table 10).
- *Matched compute* (Section B.7): all ablation configurations share the same step budget and per-GPU batch size.
- *Stability of raw-gradient PGD* (Section B.8): gradient-scale measurements of the adversarial inner loop.
- *Ablation without modality mask* (Section B.9, Table 11): the effect of applying adversarial perturbations to all modalities.
- *Prediction change rate* (Section B.10): δ_{cp} for vanilla models (Table 12) and before vs. after our training (Table 13).

Information-Theoretic Interpretation (Section C): We show that our training objective serves as a variational surrogate for a conditional Information Bottleneck objective, connecting δ_{cp} to the conditional mutual information $I(Y; X_{\mathrm{irr}} \mid X_{\mathrm{rel}})$. Radar plots of task-wise robustness are provided in Figure 6.

Hyperparameter Settings (Section D): Training configurations (Table 14), dataset statistics (Table 15), and an analysis of the number of adversarial iterations (Figure 7).

Compute Resources (Section E): Hardware, a theoretical FLOPs analysis, and measured training time across model scales (Table 16).

Limitations (Section F) and **Use of LLMs** (Section G).

Table 4: Unimodal ability evaluation on image-heavy and text-heavy tasks under perturbation. Left: Mini-ImageNet and Caltech-101; Right: OpenBookQA and MMLU. UF = Unrelated Facts, MD = Misleading Descriptions, RP = Random Pixels, RI = Real Image, FB & FW = Full Black/White Canvas. The best accuracy is marked in bold.

Model	**Method**	**Mini-ImageNet**			**Caltech-101**			**OpenBookQA**				**MMLU**			
		Orig	UF	MD	Orig	UF	MD	RP	RI	FB	FW	RP	RI	FB	FW
Qwen2.5-VL-3B [3]	Vanilla	98.9	98.5	94.9	98.8	99.0	94.4	79.9	74.6	80.0	79.7	63.5	61.1	64.0	63.6
	FFT with D^{VQA}	98.8	98.5	95.3	98.8	98.1	94.3	80.7	74.3	80.1	80.2	63.0	61.7	63.0	63.3
	FFT with D^{AUG}	98.8	98.8	98.6	99.6	99.3	99.6	**87.1**	86.7	**87.4**	**87.2**	64.8	63.9	64.8	64.7
	+ KL	98.9	98.7	99.1	99.6	99.5	**99.7**	**87.1**	86.2	86.7	86.8	**66.0**	**65.5**	**65.9**	**65.9**
	+ JS	99.1	98.8	98.3	99.6	99.4	98.0	85.0	84.2	85.1	85.1	65.6	65.1	65.5	65.5
	+ RG (σ=0.05)	99.0	99.1	98.9	99.5	99.5	99.2	86.4	**86.9**	86.9	**87.2**	64.6	64.3	64.6	64.8
	+ ADV	**99.3**	**99.3**	99.1	99.5	99.4	99.5	86.6	85.8	86.8	86.6	65.3	64.0	65.4	65.3
	Ours	**99.3**	99.2	**99.2**	**99.7**	**99.7**	99.5	86.7	86.4	86.6	86.6	64.8	64.5	65.0	65.1
Qwen2.5-VL-7B [3]	Vanilla	99.3	99.3	96.3	99.1	98.9	97.2	85.9	77.5	85.8	86.0	69.3	63.7	68.9	68.9
	FFT with D^{VQA}	99.2	99.3	96.0	99.5	99.5	95.7	86.3	82.3	86.5	86.3	69.2	67.4	69.4	69.3
	FFT with D^{AUG}	**99.6**	**99.5**	99.4	**99.7**	**99.7**	99.5	90.2	88.2	90.3	90.3	70.4	69.9	70.4	70.3
	+ KL	99.3	99.3	99.1	99.6	99.6	99.6	**92.0**	88.7	**92.2**	**92.1**	71.2	**70.7**	71.0	71.0
	+ JS	99.5	99.4	99.4	**99.7**	99.3	99.6	91.6	88.7	**92.2**	**92.1**	**71.5**	69.9	**71.5**	**71.6**
	+ RG (σ=0.05)	99.4	99.4	99.2	99.6	99.3	99.4	89.1	87.9	89.1	89.1	66.7	65.6	66.5	66.5
	+ ADV	99.4	99.3	99.2	99.6	99.5	99.6	91.7	88.8	91.8	91.8	70.4	70.0	70.4	70.4
	Ours	**99.6**	**99.5**	**99.5**	99.6	99.6	**99.7**	90.9	**89.7**	90.8	91.0	69.8	68.2	69.8	70.0
LLaVA-1.5-7B [35]	Vanilla	95.3	93.4	43.5	97.0	95.9	57.4	62.4	56.4	62.5	63.4	46.3	45.2	45.9	45.8
	FFT with D^{VQA}	94.3	92.7	41.5	96.2	94.0	46.3	61.3	55.5	62.0	62.9	46.8	45.6	47.5	47.7
	FFT with D^{AUG}	98.2	98.2	98.1	98.5	98.6	99.0	78.6	77.2	78.7	78.4	51.1	50.7	51.1	51.3
	+ KL	**99.1**	**99.0**	98.9	98.6	98.7	98.8	81.4	81.3	81.4	81.2	**52.0**	**51.8**	52.0	52.2
	+ JS	98.7	98.8	**99.0**	99.1	**99.0**	99.2	81.6	**81.5**	**81.7**	81.5	51.6	**51.8**	**52.4**	**52.4**
	+ RG (σ=0.05)	98.4	98.4	98.5	98.9	98.9	99.1	80.5	79.8	79.9	80.3	49.5	49.5	49.9	49.6
	+ ADV	98.7	98.7	98.5	98.7	98.5	98.8	81.7	81.0	81.4	80.8	50.6	50.3	50.9	50.7
	Ours	98.6	98.4	98.7	**99.3**	98.9	**99.3**	**81.8**	81.0	**81.7**	**81.7**	51.5	51.0	51.5	51.4
LLaVA-1.5-13B [35]	Vanilla	95.6	94.1	73.0	97.9	97.1	77.4	65.9	63.8	68.0	69.1	51.8	50.8	52.7	52.7
	FFT with D^{VQA}	94.6	93.9	72.0	97.8	96.5	80.2	67.5	64.2	69.1	69.3	52.4	52.2	53.1	53.3
	FFT with D^{AUG}	98.1	96.8	98.4	96.7	96.9	97.0	81.0	78.7	81.1	81.3	52.1	51.7	51.8	51.6
	+ KL	98.3	98.0	98.6	98.8	98.5	98.9	83.0	**82.6**	**83.3**	83.0	55.7	55.1	55.6	55.6
	+ JS	98.3	98.1	98.0	98.7	98.4	98.7	83.1	81.5	83.1	**83.1**	**56.7**	**56.2**	56.6	56.5
	+ RG (σ=0.05)	98.5	98.0	98.1	98.9	98.5	98.9	**83.5**	82.5	83.1	82.8	55.4	55.3	55.7	55.5
	+ ADV	**98.7**	98.2	98.6	99.0	**98.6**	99.0	82.2	**82.6**	82.6	82.8	55.6	55.4	55.6	55.5
	Ours	98.5	**98.4**	**98.7**	**99.2**	**98.6**	**99.2**	83.0	82.1	82.7	**83.1**	56.6	55.8	**56.7**	**56.7**
InstructBLIP-7B [2]	Vanilla	92.0	87.1	13.6	90.3	90.2	17.5	50.9	46.2	50.9	50.7	35.3	35.8	35.2	35.7
	FFT with D^{VQA}	95.6	86.6	16.3	98.3	91.0	23.1	49.8	45.2	49.5	50.7	40.9	40.2	41.0	41.6
	FFT with D^{AUG}	98.5	98.0	38.2	99.0	98.7	56.1	75.0	74.9	74.8	75.8	50.0	49.7	50.0	50.0
	+ KL	98.7	98.1	98.3	**99.5**	**99.0**	**99.6**	76.9	77.0	76.9	77.3	**51.3**	**50.6**	**51.3**	**51.5**
	+ JS	98.5	97.7	**98.5**	98.9	98.4	98.9	78.0	76.6	77.7	78.0	50.7	50.1	50.7	50.8
	+ RG (σ=0.05)	**98.9**	97.2	72.5	99.1	**99.0**	82.2	75.2	72.6	76.0	76.9	48.3	47.6	48.9	49.1
	+ ADV	98.7	**98.5**	32.2	**99.5**	98.9	49.2	76.8	76.8	76.5	76.3	49.3	48.4	49.5	49.4
	Ours	98.4	97.9	98.0	99.2	98.3	99.0	**79.0**	**77.3**	**79.3**	**79.0**	50.2	49.7	50.3	50.2
InstructBLIP-13B [2]	Vanilla	93.0	81.6	50.9	94.1	82.7	51.0	44.7	39.4	45.6	46.1	40.1	37.4	40.5	42.9
	FFT with D^{VQA}	95.6	85.8	8.0	97.0	87.5	11.6	58.6	55.4	59.7	60.6	43.7	42.8	43.6	44.1
	FFT with D^{AUG}	98.4	98.2	9.3	99.2	98.8	13.8	82.0	80.4	81.2	81.2	52.1	51.3	52.4	53.0
	+ KL	98.5	**98.3**	98.7	99.1	99.2	**99.5**	82.5	81.4	82.1	82.9	**53.4**	**52.5**	**53.4**	**53.4**
	+ JS	98.7	98.1	**98.9**	99.3	99.2	**99.5**	**83.5**	**83.1**	83.1	**83.3**	52.8	52.2	53.2	53.3
	+ RG (σ=0.05)	98.4	97.8	87.0	**99.4**	**99.3**	94.4	80.0	76.6	79.6	80.4	50.9	50.0	51.4	51.8
	+ ADV	98.6	98.0	80.9	98.7	98.6	99.1	79.8	79.0	80.9	80.9	51.3	50.7	51.4	52.4
	Ours	**98.7**	97.9	98.0	98.7	98.7	98.8	83.2	81.2	**83.8**	83.0	52.2	51.6	52.3	**53.4**

Table 5: Evaluation of unimodal and multimodal tasks across different ablation study settings. For unimodal datasets, we report accuracy on the original setting (Orig) and the worst-performing perturbation (Perturbed). For VQA datasets, we report accuracy on the original setting. The best accuracy is marked in bold.

Model	Method	Mini-ImageNet		Caltech-101		OpenBookQA		MMLU		VQA Overall
		Orig	Perturbed	Orig	Perturbed	Orig	Perturbed	Orig	Perturbed	Accuracy
3B Multimodal Models										
Qwen2.5-VL-3B [3]	Vanilla	98.9	94.9	98.8	94.4	79.9	74.6	63.5	61.1	72.5
	FFT with D^{VQA}	98.8	95.3	98.8	94.3	80.7	74.3	63.0	61.7	75.5
	FFT with D^{AUG}	98.8	98.6	99.6	99.3	**87.1**	86.7	64.8	63.9	76.2
	+ KL	98.9	98.7	99.6	**99.5**	**87.1**	86.2	**66.0**	**65.5**	75.9
	+ JS	99.1	98.3	99.6	98.0	85.0	84.2	65.6	65.1	76.2
	+ RG	99.0	98.9	99.5	99.2	86.4	**86.9**	64.6	64.3	76.0
	+ ADV	**99.3**	99.1	99.5	99.4	86.6	85.8	65.3	64.0	76.1
	Ours	**99.3**	**99.2**	**99.7**	**99.5**	86.7	86.4	64.8	64.5	**76.4**
7B Multimodal Models										
LLaVA-1.5-7B [35]	Vanilla	95.3	43.5	97.0	57.4	62.4	56.4	46.3	45.2	63.8
	FFT with D^{VQA}	94.3	41.5	96.2	46.3	61.3	55.5	46.8	45.6	64.0
	FFT with D^{AUG}	98.2	98.1	98.5	98.6	78.6	77.2	51.1	50.7	65.6
	+ KL	**99.1**	**98.9**	98.6	98.7	81.4	81.2	**52.0**	**51.8**	66.3
	+ JS	98.7	98.8	99.1	**99.0**	81.6	**81.5**	51.6	**51.8**	65.6
	+ RG	98.4	98.4	98.9	98.9	80.5	79.8	49.5	49.5	65.7
	+ ADV	98.7	98.5	98.7	98.5	81.7	80.8	50.6	50.3	65.7
	Ours	98.6	98.4	**99.3**	98.9	**81.8**	81.0	51.5	51.0	**66.8**
InstructBLIP-7B [2]	Vanilla	92.0	13.6	90.3	17.5	50.9	46.2	35.3	35.2	56.4
	FFT with D^{VQA}	95.6	16.3	98.3	23.1	49.8	45.2	40.9	40.2	60.4
	FFT with D^{AUG}	98.5	38.2	99.0	56.1	75.0	74.8	50.0	49.7	63.9
	+ KL	98.7	**98.1**	**99.5**	**99.0**	76.9	76.9	**51.3**	**50.6**	63.9
	+ JS	98.5	97.7	98.9	98.4	78.0	76.6	50.7	50.1	64.1
	+ RG	**98.9**	72.5	99.1	82.2	75.2	72.6	48.3	47.6	64.2
	+ ADV	98.7	32.2	**99.5**	49.2	76.8	76.3	49.3	48.4	64.0
	Ours	98.4	97.9	99.2	98.3	**79.0**	**77.3**	50.2	49.7	**64.8**
Qwen2.5-VL-7B [3]	Vanilla	99.3	96.3	99.1	97.2	85.9	77.5	69.3	63.7	80.3
	FFT with D^{VQA}	99.2	96.0	99.5	95.7	86.3	82.3	69.2	67.4	79.5
	FFT with D^{AUG}	**99.6**	99.4	**99.7**	99.5	90.2	88.2	70.4	69.9	79.9
	+ KL	99.3	99.1	99.6	**99.6**	**92.0**	88.7	71.2	**70.7**	80.6
	+ JS	99.5	99.4	**99.7**	99.3	91.6	88.7	**71.5**	69.9	80.3
	+ RG	99.4	99.2	99.6	99.3	89.1	87.9	66.7	65.6	78.0
	+ ADV	99.4	99.2	99.6	99.5	91.7	88.8	70.4	70.0	79.9
	Ours	**99.6**	**99.5**	99.6	**99.6**	90.9	**89.7**	69.8	68.2	**80.9**
13B Multimodal Models										
LLaVA-1.5-13B [35]	Vanilla	95.6	73.0	97.9	77.4	65.9	63.8	51.8	50.8	66.2
	FFT with D^{VQA}	94.6	72.0	97.8	80.2	67.5	64.2	52.4	52.2	65.8
	FFT with D^{AUG}	98.1	96.8	96.7	96.9	81.0	78.7	52.1	51.6	67.1
	+ KL	98.3	98.0	98.8	98.5	83.0	**82.6**	55.7	55.1	68.0
	+ JS	98.3	98.0	98.7	98.4	83.1	81.5	**56.7**	**56.2**	68.6
	+ RG	98.5	98.0	98.9	98.5	**83.5**	82.5	55.4	55.3	66.9
	+ ADV	**98.7**	98.2	99.0	**98.6**	82.2	**82.6**	55.6	55.4	67.5
	Ours	98.5	**98.4**	**99.2**	**98.6**	83.0	82.1	56.6	55.8	**69.7**
InstructBLIP-13B [2]	Vanilla	93.0	50.9	94.1	51.0	44.7	39.4	40.1	37.4	65.8
	FFT with D^{VQA}	95.6	8.0	97.0	11.6	58.6	55.4	43.7	42.8	59.4
	FFT with D^{AUG}	98.4	9.3	99.2	13.8	82.0	80.4	52.1	51.3	65.9
	+ KL	98.5	**98.3**	99.1	**99.2**	82.5	81.4	**53.4**	**52.5**	66.2
	+ JS	**98.7**	98.1	99.3	**99.2**	**83.5**	**83.1**	52.8	52.2	66.5
	+ RG	98.4	87.0	**99.4**	94.4	80.0	76.6	50.9	50.0	66.3
	+ ADV	98.6	80.9	98.7	98.6	79.8	79.0	51.3	50.7	66.4
	Ours	**98.7**	97.9	98.7	98.7	83.2	81.2	52.2	51.6	**67.7**

Table 6: Detailed multimodal reasoning accuracy (%) on multiple-choice VQA datasets across different ablation study settings with extra models: Qwen2.5-vl-7B, Instructblip-Vicuna-13B. The best accuracy is marked in bold. Overall performance is computed as a weighted average across datasets, with weights proportional to each dataset's test size.

Model	Method	ScienceQA-IMG	MM-Bench-EN	Seed-Bench-IMG	VQA Overall
Qwen2.5-VL Models [3]					
Qwen2.5-VL-3B [3]	Vanilla	63.0 (100%)	81.8 (100%)	72.3 (100%)	72.5 (100%)
	FFT with D^{VQA}	72.2 (↑14.6%)	82.0 (↑0.2%)	74.7 (↑3.3%)	75.5 (↑3.9%)
	FFT with D^{AUG}	73.6 (↑16.8%)	81.7 (↓0.1%)	75.3 (↑4.1%)	76.2 (↑5.1%)
	+ KL	72.8 (↑15.6%)	80.9 (↓1.1%)	74.9 (↑3.6%)	75.9 (↑4.7%)
	+ JS	73.4 (↑16.5%)	82.0 (↑0.2%)	75.2 (↑4.0%)	76.2 (↑5.1%)
	+ RG	73.4 (↑16.5%)	81.8 (↑0.0%)	75.0 (↑3.7%)	76.0 (↑4.8%)
	+ ADV	73.1 (↑16.0%)	81.7 (↓0.1%)	75.3 (↑4.1%)	76.1 (↑5.0%)
	Ours	73.8 (↑17.1%)	82.1 (↑0.1 %)	75.5 (↑4.4%)	**76.4 (↑5.4%)**
Qwen2.5-VL-7B [3]	Vanilla	85.1 (100%)	86.6 (100%)	77.0 (100%)	80.3 (100%)
	FFT with D^{VQA}	81.4 (↓4.3%)	86.4 (↓0.2%)	76.8 (↓0.3%)	79.5 (↓1.0%)
	FFT with D^{AUG}	83.2 (↓2.2%)	86.1 (↓0.6%)	77.0 (-0.0%)	79.9 (↓0.5%)
	+ KL	86.0 (↑1.1%)	86.7 (↑0.1%)	77.2 (↑0.3%)	80.6 (↑0.4%)
	+ JS	85.5 (↑0.5%)	85.9 (↓0.8%)	77.0 (-0.0%)	80.3 (-0.0%)
	+ RG	81.6 (↓4.1%)	82.9 (↓4.2%)	75.5 (↓1.9%)	78.0 (↓2.8%)
	+ ADV	85.2 (↑0.1%)	85.8 (↓0.9%)	76.6 (↓0.5%)	79.9 (↓0.5%)
	Ours	83.9 (↓1.4%)	86.4 (↓0.2%)	78.3 (↑1.7%)	**80.9 (↑0.7%)**
Instructblip-Vicuna Models [2]					
InstructBLIP-7B [2]	Vanilla	52.1 (100%)	65.5 (100%)	54.8 (100%)	56.4 (100%)
	FFT with D^{VQA}	61.8 (↑18.6%)	68.4 (↑4.4%)	57.6 (↑5.1%)	60.4 (↑7.1%)
	FFT with D^{AUG}	65.4 (↑25.5%)	71.1 (↑8.5%)	61.2 (↑11.7%)	63.9 (↑13.3%)
	+ KL	65.9 (↑26.5%)	71.7 (↑9.5%)	60.9 (↑11.1%)	63.9 (↑13.3%)
	+ JS	66.9 (↑28.4%)	72.3 (↑10.4%)	60.7 (↑10.8%)	64.1 (↑13.7%)
	+ RG	62.8 (↑20.5%)	70.9 (↑8.2%)	62.5 (↑14.0%)	64.2 (↑13.8%)
	+ ADV	66.0 (↑26.7%)	69.0 (↑5.3%)	61.8 (↑12.8%)	64.0 (↑13.5%)
	Ours	64.0 (↑22.8%)	71.4 (↑9.0%)	63.0 (↑14.9%)	**64.8 (↑14.9%)**
InstructBLIP-13B [2]	Vanilla	65.8 (100%)	72.1 (100%)	63.8 (100%)	65.8 (100%)
	FFT with D^{VQA}	61.7 (↓6.2%)	68.5 (↓5.0%)	56.0 (↓12.2%)	59.4 (↓9.7%)
	FFT with D^{AUG}	66.6 (↑1.2%)	71.5 (↓0.8%)	64.0 (↑0.3%)	65.9 (↑0.2%)
	+ KL	67.2 (↑2.2%)	72.6 (↑0.7%)	64.0 (↑0.3%)	66.2 (↑0.6%)
	+ JS	67.9 (↑3.2%)	72.7 (↑0.8%)	64.2 (↑0.6%)	66.5 (↑1.1%)
	+ RG	65.4 (↓0.6%)	73.5 (↑1.9%)	64.3 (↑0.7%)	66.3 (↑0.8%)
	+ ADV	66.1 (↑0.5%)	74.0 (↑2.6%)	64.1 (↑0.5%)	66.4 (↑0.9%)
	Ours	68.2 (↑3.6%)	74.4 (↑3.2%)	65.5 (↑1.0%)	**67.7 (↑2.8%)**
LLaVA1.5 Models [35]					
LLaVA-1.5-7B [35]	Vanilla	64.5 (100%)	64.3 (100%)	63.4 (100%)	63.8 (100%)
	FFT with D^{VQA}	61.6 (↓4.5%)	71.4 (↑11.0%)	62.4 (↓1.6%)	64.0 (↑0.3%)
	FFT with D^{AUG}	65.4 (↑1.4%)	71.1 (↑10.6%)	63.9 (↑0.8%)	65.6 (↑2.8%)
	+ KL	65.9 (↑2.2%)	72.5 (↑12.8%)	64.5 (↑1.7%)	66.3 (↑3.9%)
	+ JS	63.8 (↓1.1%)	73.5 (↑14.3%)	63.7 (↑0.5%)	65.6 (↑2.8%)
	+ RG	66.3 (↑2.8%)	71.8 (↑11.7%)	63.7 (↑0.5%)	65.7 (↑3.0%)
	+ ADV	66.7 (↑3.4%)	71.4 (↑11.0%)	63.6 (↑0.3%)	65.7 (↑3.0%)
	Ours	67.8 (↑5.1%)	73.1 (↑13.7%)	64.6 (↑1.9%)	**66.8 (↑4.7%)**
LLaVA-1.5-13B [35]	Vanilla	65.8 (100%)	72.1 (100%)	64.5 (100%)	66.2 (100%)
	FFT with D^{VQA}	60.8 (↓7.6%)	73.6 (↑2.1%)	64.9 (↑0.6%)	65.8 (↓0.6%)
	FFT with D^{AUG}	63.5 (↓3.5%)	75.0 (↑4.0%)	65.7 (↑1.9%)	67.1 (↑1.4%)
	+ KL	62.5 (↓5.0%)	74.6 (↑3.5%)	67.6 (↑4.8%)	68.0 (↑2.7%)
	+ JS	65.8 (↑0.0%)	74.7 (↑3.6%)	67.6 (↑4.8%)	68.6 (↑3.6%)
	+ RG	58.3 (↓11.4%)	73.5 (↑1.9%)	67.4 (↑4.5%)	66.9 (↑1.1%)
	+ ADV	57.6 (↓12.5%)	74.2 (↑2.9%)	68.3 (↑5.9%)	67.5 (↑2.0%)
	Ours	65.3 (↓0.8%)	73.6 (↑2.1%)	69.7 (↑8.1%)	**68.4 (↑3.3%)**

Note. While Qwen2.5-VL was originally instruction-tuned with proprietary in-house data [3], our reproduced version uses only publicly available LLaVA instruction-tuning data. Even under this constraint and without access to VQA-specific tuning samples, our models achieve comparable or even better performance across all VQA datasets—highlighting the robustness and effectiveness of our proposed perturbation-consistent fine-tuning strategy.

B Detailed Experimental Results

Please see Table 4, Table 5, Figure 6, Table 7, Table 6, Table 10, Table 9 and Table 8 for more details. In table Table 6: We compare the pretrained models with the following strategies: *FFT with* D^{VQA} (standard finetuning on VQA data), *FFT with* D^{AUG} (supervised finetuning on mixed multi-task datasets with heuristic perturbations), *FFT+KL/JS* (adding consistency regularization on KL or JS divergence), *FFT+RG* (injecting random Gaussian noise into token embeddings), *FFT+ADV* (FFT with heuristic & adversarial perturbations), and *Ours* (combining both perturbation-based data augmentation and consistency regularization).

B.1 Finetuning Strategies

In our adaptation of InstructBLIP-Vicuna-7B, we choose to freeze the Q-Former and only fine-tune the language model and the projection layer. This decision is grounded in the nature of the Q-Former as a highly task-specific visual query encoder, originally pre-trained on VQA-style datasets where fine-grained and semantically aligned image-text pairs dominate.

However, in our setting, we deliberately introduce perturbations to the input modalities (e.g., injecting unrelated or misleading text/image content), which breaks the expected alignment structure. We observe that training the Q-Former under such noisy supervision leads to unstable representations and overfitting to spurious modality correlations. In contrast, freezing the Q-Former allows us to preserve its original strong visual grounding capabilities, while letting the downstream language model learn to filter or suppress misleading signals introduced during training.

This alternative tuning strategy enhances robustness under modality interference and aligns with our overall goal of improving cross-modal competency in MLLMs under perturbed conditions.

Table 7: Performance (%) of Vanilla models under modality interference across four datasets. We show accuracy under clean (origin) and various perturbations: Left: Mini-ImageNet and Caltech-101; Right: OpenBookQA and MMLU. UF = Unrelated Facts, MD = Misleading Descriptions, RP = Random Pixels, RI = Real Image, FB & FW = Full Black/White Canvas. (The results are averaged on multiple runs with standard deviation < 0.2)

Model	Mini-ImageNet			Caltech-101			Open-Book QA				MMLU			
	Orig	UF	MD	Orig	UF	MD	RP	RI	FB	FW	RP	RI	FB	FW
InternVL2-2B [51]	91.9	88.6	25.5	94.6	91.3	33.2	46.3	36.2	45.2	45.3	39.1	36.9	39.3	39.2
LLaVA-1.5-7B [35]	95.3	93.4	43.5	97.0	95.9	57.4	62.4	56.4	62.5	63.4	46.3	45.2	45.9	45.8
LLaVA-Next-7B [52]	93.4	90.0	28.5	97.0	93.4	31.6	54.6	52.9	55.9	55.5	45.9	45.0	45.9	45.8
LLaVA-Next-34B [52]	98.0	96.3	90.2	99.1	97.3	93.6	87.7	85.7	88.4	88.0	71.5	70.9	71.5	71.5
LLaVA-Next-72B [52]	97.7	97.8	83.0	98.9	98.8	91.2	88.2	86.2	89.1	88.9	73.6	72.9	73.9	73.9
LLaVA-Next-110B [52]	98.4	98.3	93.3	98.4	98.3	93.1	89.5	89.2	89.9	89.7	73.5	73.0	73.9	73.9
LLaVA-1.5-13B [35]	95.6	94.1	73.0	97.9	97.1	77.4	65.9	63.8	68.0	69.1	51.8	50.8	52.7	52.7
InstructBLIP-7B [2]	92.0	87.1	13.6	90.3	90.2	17.5	50.9	46.2	50.9	50.7	35.3	35.8	35.2	35.7
InstructBLIP-13B [2]	93.0	81.6	50.9	94.1	82.7	51.0	44.7	39.4	45.6	46.1	40.1	37.4	40.5	42.9
QwenVL2-2B [53]	98.7	98.6	75.8	99.1	99.1	66.8	59.6	54.7	63.4	63.8	49.5	44.6	49.7	50.0
QwenVL2.5-3B [3]	98.9	98.5	94.9	98.8	99.0	94.4	79.9	74.6	80.0	79.7	63.5	61.1	64.0	63.6
QwenVL3-4B [54]	99.3	98.9	96.5	99.3	99.2	98.0	84.1	82.9	86.2	86.0	71.8	70.9	72.4	72.4
QwenVL2-7B [53]	99.0	99.1	96.3	99.6	99.5	97.6	82.5	80.9	83.7	83.2	66.9	65.3	67.7	67.8
QwenVL2.5-7B [3]	99.3	99.3	96.3	99.1	98.9	97.2	85.9	77.5	85.8	86.0	69.3	63.7	68.9	68.9
QwenVL3-8B [54]	99.3	98.8	97.2	99.8	99.7	98.1	87.0	85.7	87.5	88.1	75.6	75.0	76.2	76.3

B.2 Perturbation-based Evaluation Experiment Results

We conduct a controlled perturbation-based evaluation across various MLLMs, as shown in Table 7. Our results reveal that both vision and language tasks are vulnerable to cross-modal interference. In vision classification tasks, misleading textual descriptions (e.g., text contradicting image content) lead to severe performance drops. Conversely, for language tasks such as OpenBookQA and MMLU, irrelevant visual inputs—particularly semantically unrelated real images—also degrade performance.

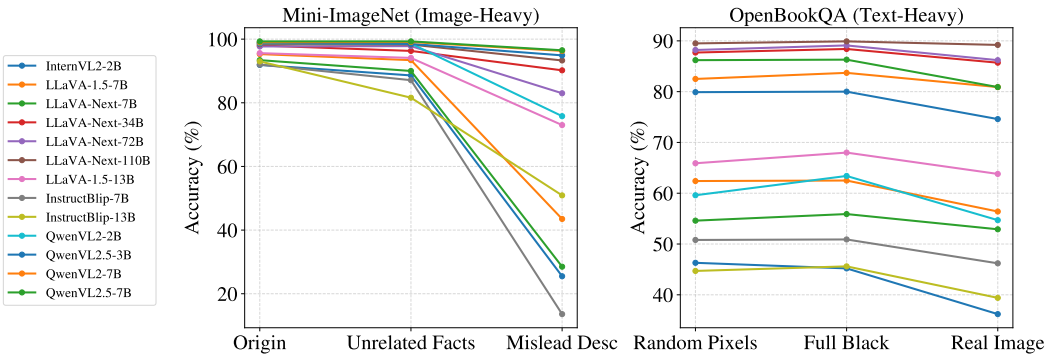


Figure 5: Performance degradation under irrelevant perturbations reveals modality interference in MLLMs. Left: Mini-ImageNet (image-heavy) with Original input, Unrelated Facts, and Misleading Descriptions. Right: OpenBookQA (text-heavy) with Random Pixels, Full Black Canvas, and Irrelevant Real Images. Misleading descriptions most severely affect image-heavy tasks, while irrelevant real images cause the largest drop in text-heavy reasoning.

A consistent trend is that larger models exhibit greater robustness. Models like LLaVA-1.5-13B and QwenVL2.5-7B maintain high accuracy across all perturbation types, indicating improved modality disentanglement and reduced sensitivity to spurious correlations. Nonetheless, performance still degrades relative to clean inputs, highlighting that interference effects remain non-negligible even in stronger models.

We further observe a clear scaling trend within the LLaVA-Next family. As model size increases from 7B to 34B, 72B, and 110B, performance under perturbations steadily improves, reflecting stronger representation power and enhanced robustness to spurious cues. For instance, LLaVA-Next-7B achieves 90.3% on Mini-ImageNet (Orig) but drops to 28.5% with misleading descriptions, whereas LLaVA-Next-110B maintains 98.4% and 93.3% under the same conditions. Similarly, on MMLU, accuracy under irrelevant real images increases from 45.9% (7B) to 73.6% (72B). **These results confirm that scaling up helps mitigate modality interference. However, such gains come at substantial computational and resource costs, and the improvements remain incremental relative to the clean–perturbed gap.** This underscores that scaling alone is insufficient, and more targeted interventions—such as our proposed framework—are necessary for robust cross-modal reasoning.

B.3 Discussion on KL&JS use for Consistency Regularization

Although both KL and JS divergence serve as effective objectives for consistency regularization, we find that JS consistently achieves slightly better results across most settings. Specifically, in both unimodal tasks (e.g., Mini-ImageNet, MMLU) and multimodal reasoning benchmarks (e.g., ScienceQA, SeedBench), JS-regularized models consistently outperform their KL counterparts by a small but observable margin. This trend holds across different model backbones and training configurations, including our final unified method (see "Ours" rows in Table 4 and 6). This suggests a marginal advantage of JS regularization in enhancing model robustness.

B.4 Evaluating Chain-of-Thought Prompting for Modality Interference Mitigation

To further investigate the potential of prompt-based methods in mitigating modality interference, we conduct additional experiments using Chain-of-Thought [33] style prompting. This approach aims to encourage structured reasoning by guiding the model through an explicit reasoning process before producing its final answer.

Specifically, we prepend the following CoT prompt to each input question:

```
Let's think step by step:
1. What information does the image provide?
2. What is the question asking?
3. Are there any misleading parts?
```

```
4. Now give your final answer. Only write the final answer on a
separate line like: "Answer: B"
```

B.5 Evaluation on Free-Form VQA

To further assess the generalizability of our method beyond multiple-choice VQA tasks, we evaluate on TextVQA [55], a free-form generative visual question answering dataset that requires reasoning over both textual and visual content in natural images. We follow the benchmark adopted by LLaVA [1], which evaluates a model's ability to both recognize textual characters within images and effectively handle noisy outputs generated by OCR systems.

Following standard evaluation protocols and existing MLLM baselines (e.g., LLaVA, Qwen2.5-VL, InstructBLIP), we report model performance averaged over multiple runs (standard deviation < 0.4). Results are presented in Table 8.

Our method achieves consistent improvements across most model families, indicating its effectiveness not only in MCQA scenarios but also in open-ended multimodal reasoning settings.

Table 8: Accuracy (%) on the TextVQA dataset across different model families.

Setting	Model	TextVQA
Vanilla	InstructBLIP-Vicuna-7B	19.7
Ours	InstructBLIP-Vicuna-7B	**32.4**
Vanilla	LLaVA-1.5-7B	49.8
Ours	LLaVA-1.5-7B	**51.2**
Vanilla	Qwen2.5-VL-7B	81.4
Ours	Qwen2.5-VL-7B	**84.8**

B.6 Improving Generalizability through Adversarial Training

To assess the generalization benefits of adversarial training, we evaluate model robustness under two types of out-of-distribution (OOD) perturbations at test time:

- **Document OCR noise:** Real-world noisy OCR snippets are sampled from the FUNSD dataset [47] and inserted as irrelevant textual distractors into image-heavy visual classification tasks (Mini-ImageNet, Caltech-101).
- **Unrelated screenshots:** Unrelated UI screenshots are drawn from the RICO dataset [48] and added as visual distractors to text-heavy reasoning tasks (OpenBookQA, MMLU).

Beyond these seen datasets, we further evaluate generalization on two *completely unseen datasets* at test time: **Food-101** and **ARC Challenge**. Food-101 serves as an unseen image-heavy classification benchmark, where models are exposed to diverse textual perturbations, including unrelated descriptions, misleading class cues, and OCR-style noise. ARC Challenge represents an unseen text-heavy reasoning task, where various visual distractors—such as random pixels, unrelated real images, full-black/white images, and UI screenshots—are introduced to probe robustness against spurious visual signals.

Each experiment is repeated multiple times, and we report average accuracy across runs (standard deviation < 0.1). Results on seen datasets are presented in Table Table 9, while results on unseen datasets are summarized in Table 10. Across all model families and task types, adversarial training consistently improves robustness to both unseen perturbation types and unseen datasets. In contrast, heuristic data augmentation alone often fails to generalize beyond the perturbation patterns observed during training and can even degrade performance under novel noise.

Overall, these results demonstrate that adversarial training substantially enhances out-of-domain generalization with only modest additional training overhead. This improved robustness to distribution shifts is crucial for reliable deployment of multimodal models in real-world, noisy environments.

B.7 Matched compute across ablation rows.

All ablation configurations in Table 6 are trained for the same number of steps with the same per-GPU batch size; the rows differ only in the *composition* of each training batch and in the training objective

Table 9: Accuracy (%) on original and perturbed inputs. OCR snippets are inserted into image classification tasks, and RICO screenshots are added to VQA tasks.

Setting	Model	Mini-ImageNet		Caltech-101		OpenBookQA		MMLU	
		Origin	OCR	Origin	OCR	RandPixels	Screenshot	RandPixels	Screenshot
vanilla	instructblip-vicuna-7b	92.0	81.4	90.3	83.6	50.9	40.0	35.3	34.9
SFT	instructblip-vicuna-7b	98.5	95.7	99.0	98.4	75.0	74.4	50.0	49.2
SFT + ADV	instructblip-vicuna-7b	98.7	97.5	99.5	98.8	76.8	76.2	49.3	49.3
Ours	instructblip-vicuna-7b	98.4	97.3	99.2	99.0	79.0	77.8	50.2	49.2
vanilla	llava-1.5-7b	95.3	88.9	97.0	92.8	62.4	50.2	46.3	44.8
SFT	llava-1.5-7b	98.2	98.0	98.5	98.0	78.6	78.0	51.1	50.6
SFT + ADV	llava-1.5-7b	98.7	98.2	98.7	98.4	81.7	81.3	50.6	51.3
Ours	llava-1.5-7b	98.6	98.2	99.3	99.0	81.8	81.2	51.5	51.3
vanilla	qwen2.5-vl-7b	99.3	99.2	99.1	99.2	85.9	69.3	69.3	57.0
SFT	qwen2.5-vl-7b	99.6	99.5	99.7	99.3	90.2	85.8	70.4	63.7
SFT + ADV	qwen2.5-vl-7b	99.4	99.5	99.6	99.6	91.7	91.5	70.4	69.7
Ours	qwen2.5-vl-7b	99.6	99.5	99.6	99.5	90.9	90.7	69.8	69.7

Table 10: Accuracy (%) on Food-101 (image-heavy) and ARC Challenge (text-heavy) under diverse test-time perturbations. For Food-101, we add different types of textual distractors (unrelated, misleading, and OCR-style noise). For ARC Challenge, we use various visual distractors (random pixels, unrelated real images, full-black/white images, and UI screenshots).

Setting	Model	Food-101				ARC Challenge				
		Origin	Unrelated Text	Mislead Text	OCR Text	RandPixels	Real Image	Full Black	Full White	Screenshot
vanilla	qwen2.5-vl-7b	97.1	96.8	93.3	96.7	85.6	83.7	85.9	85.9	79.6
$+D^{\text{AUG}}$	qwen2.5-vl-7b	97.1	96.6	94.1	96.5	85.0	82.9	84.9	84.9	80.2
Ours	qwen2.5-vl-7b	97.5	97.5	97.6	97.4	84.1	82.7	84.2	84.0	82.6
vanilla	llava-1.5-7b	90.2	86.3	31.0	83.6	53.4	51.6	54.5	54.0	51.8
$+D^{\text{AUG}}$	llava-1.5-7b	91.3	86.5	40.2	83.5	55.7	52.0	56.0	56.0	53.8
Ours	llava-1.5-7b	92.1	89.7	69.0	89.5	62.8	62.0	62.6	62.8	62.6
vanilla	instructblip-vicuna-7b	67.9	20.1	18.7	48.0	33.7	35.8	33.7	31.2	36.5
$+D^{\text{AUG}}$	instructblip-vicuna-7b	55.5	41.4	38.4	56.5	27.3	28.2	27.6	27.2	30.7
Ours	instructblip-vicuna-7b	95.6	89.8	93.7	88.1	57.6	56.4	58.8	59.6	57.0

applied. Specifically, D^{VQA} fills each batch with VQA samples, whereas D^{AUG} fills each batch with a fixed $50/25/25$ mixture of VQA, image-heavy, and text-heavy samples (with perturbed variants sharing the same batch budget rather than being added on top). The $+$KL, $+$ADV, and Ours rows further share the same D^{AUG} data mixture and step budget, and introduce only the consistency regularizer or adversarial perturbation as additional objectives. Reported gains across rows therefore reflect changes in data composition or training objective at matched compute, not additional data or training time.

B.8 Gradient scale and stability of raw-gradient PGD.

We instrument the exact PGD loop used in training (α=0.1, ϵ=10^{-3}, T=2) on Qwen3-VL-4B over 200 training batches (2,800 sample-steps). Raw gradient scales differ substantially across modalities: the median per-sample maximum gradient coordinate is 9.3×10^{-4} on visual tokens and 2.8×10^{-1} on textual tokens. Nevertheless, no perturbation leaves the feasible region and no loss divergence occurs. On image-heavy samples, the unconstrained update exceeds the perturbation radius by $28\times$ at the median (up to $142\times$), and the projection clips these coordinates as intended. Only about 1% of perturbed coordinates finish at the $\pm\epsilon$ boundary, compared with all coordinates with a nonzero gradient under a sign-based update at the same step size. Projection therefore provides stability, while the raw gradient preserves coordinate-wise sensitivity; we do not find gradient normalization necessary.

B.9 Ablation without Modality Mask

To further examine the role of modality-aware masking, we implement a *No Mask* variant, where adversarial perturbations are applied uniformly to all modalities across all datasets, including VQA tasks.

Table 11 reports the results on representative image-heavy and text-heavy benchmarks. We observe three key findings. First, training without a modality mask still partially mitigates modality inter-

ference. This is because the combination of consistency regularization and heuristic perturbations already encourages the model to learn certain modality-invariant features, and adversarial training itself improves generic robustness (e.g., on OpenBookQA with screenshot distractors, accuracy improves from 50.2% to 62.7%).

Second, the modality mask is critical for out-of-distribution generalization. Without masking, performance collapses on unseen perturbations, particularly under strong visual distractions. For example, on OpenBookQA with screenshot perturbations, the No Mask variant underperforms the masked version by a large margin (62.7% vs. 81.8%).

Third, modality-aware masking is essential to preserve multimodal semantics during VQA training. When adversarial noise is injected into all token embeddings indiscriminately, the model fails to maintain coherent cross-modal representations, resulting in a notable degradation in overall VQA performance (61.5% vs. 66.8%).

Together, these results demonstrate that while adversarial training alone improves robustness, modality-aware masking plays a crucial role in preventing semantic destruction and enabling reliable generalization under unseen distribution shifts.

Table 11: Ablation study without modality-aware masking. The *No Mask* variant applies adversarial perturbations to all modalities across all datasets. Results are reported on image-heavy (Caltech-101) and text-heavy (OpenBookQA) benchmarks, together with overall VQA accuracy.

Model	**Caltech-101**			**OpenBookQA**			**VQA**
	Orig	Mislead Text	OCR Text	RandPixels	Real Image	Screenshot	Overall
LLaVA-1.5-7B (Vanilla)	97.0	57.4	92.8	62.4	56.4	50.2	63.8
LLaVA-1.5-7B (Ours, No Mask)	98.1	96.3	93.5	79.5	78.1	62.7	61.5
LLaVA-1.5-7B (Ours)	**99.3**	**98.9**	**99.0**	**81.8**	**81.0**	**81.8**	**66.8**

B.10 Prediction Change Rate (δ_{cp})

We report δ_{cp} here, computed as the fraction of examples whose predicted option differs between the reference and the perturbed input, over examples with a parseable prediction in both settings. The reference is the clean question for image-heavy tasks and the random-pixel canvas (RP), a semantically empty image, for text-heavy tasks.

Accuracy is a lossy summary of δ_{cp}: correct-to-incorrect and incorrect-to-correct changes cancel in aggregate accuracy but both count toward δ_{cp}. Table 12 shows that on text-heavy tasks δ_{cp} is several times larger than the accuracy change, stays substantial even when accuracy is unchanged or improves (e.g., under FB), and remains non-zero for the newest and largest models. For example, attaching an information-free black canvas raises QwenVL3-8B's OpenBookQA accuracy by 0.5 points while changing 5.6% of its answers. Table 13 shows that our framework reduces δ_{cp} under every reported intervention across all five backbones.

Table 12: Accuracy change Δ (reference $-$ perturbed, %) and prediction change rate δ_{cp} (%, bold) of vanilla models. MD = Misleading Descriptions, RI = Real Image, FB = Full Black canvas; reference settings are described in Appendix B.10.

Model	Caltech-101 MD		Mini-ImageNet MD		OpenBookQA RI		OpenBookQA FB		MMLU RI		MMLU FB	
	Δ	δ_{cp}	Δ	δ_{cp}	Δ	δ_{cp}	Δ	δ_{cp}	Δ	δ_{cp}	Δ	δ_{cp}
LLaVA-1.5-7B	39.6	**42.0**	51.8	**55.0**	6.0	**21.5**	-0.1	**10.2**	-0.1	**13.6**	–	–
LLaVA-1.5-13B	20.6	**22.5**	22.6	**25.9**	2.1	**16.7**	-2.1	**10.0**	1.0	**9.1**	-0.9	**5.5**
LLaVA-Next-7B	65.4	**67.5**	64.9	**69.5**	1.7	**14.5**	-1.3	**9.0**	1.0	**11.4**	0.0	**6.8**
QwenVL2-2B	32.3	**33.0**	22.7	**23.9**	-0.1	**24.0**	-3.8	**14.9**	0.1	**13.9**	-0.2	**15.7**
QwenVL2-7B	2.0	**2.0**	2.4	**2.6**	1.6	**12.1**	-1.2	**5.6**	0.6	**8.9**	-0.8	**5.3**
QwenVL2.5-3B	4.4	**5.4**	4.0	**4.8**	5.3	**12.9**	-0.1	**3.1**	2.4	**11.2**	-0.5	**3.8**
QwenVL2.5-7B	2.4	**2.9**	2.9	**3.3**	6.0	**11.1**	0.2	**1.6**	2.7	**11.1**	0.4	**2.9**
QwenVL3-4B	1.1	**1.4**	2.9	**3.0**	1.2	**9.0**	-2.1	**5.1**	0.9	**6.6**	-0.5	**3.9**
QwenVL3-8B	1.3	**1.3**	1.7	**2.0**	0.6	**7.0**	-0.5	**5.6**	0.5	**5.3**	-0.5	**4.4**

Table 13: Prediction change rate δ_{cp} (%, lower is better) of vanilla models and models trained with our framework, under the same interventions as Table 12.

Model	Setting	Caltech-101 MD	Mini-ImageNet MD	OpenBookQA RI	OpenBookQA FB	MMLU RI	MMLU FB
QwenVL2.5-3B	Vanilla	5.4	4.8	12.9	3.1	11.2	3.8
	+ Ours	**0.0**	**0.7**	**3.9**	**0.6**	**4.2**	**1.1**
QwenVL2.5-7B	Vanilla	2.9	3.3	11.1	1.6	11.1	2.9
	+ Ours	**0.2**	**0.3**	**4.6**	**0.4**	**6.9**	**1.3**
QwenVL3-4B	Vanilla	1.4	3.0	9.0	5.1	6.6	3.9
	+ Ours	**0.2**	**0.9**	**4.2**	**2.8**	**4.7**	**2.8**

C Information-Theoretic Interpretation of the Framework

While our framework is grounded in causal intervention (Section 3), it admits a complementary interpretation through the lens of the *Information Bottleneck* (IB) principle [56]. This section formalizes that connection and shows that our training objective serves as a tractable variational surrogate for a conditional IB objective targeting modality interference.

C.1 Setup and Notation

Let X_{rel} denote the task-relevant modality input (e.g., the image in image-heavy tasks, or the text in text-heavy tasks) and X_{irr} denote the task-irrelevant modality. Let Y denote the model's prediction and A the ground-truth label. We write $Z = f_\theta(X_{\mathrm{rel}}, X_{\mathrm{irr}})$ for the fused multimodal representation produced by an MLLM with parameters θ, and $p_\theta(Y \mid X_{\mathrm{rel}}, X_{\mathrm{irr}})$ for the induced predictive distribution.

C.2 Modality Interference as Excess Mutual Information

Under ideal cross-modality competency, the prediction Y should depend on X_{irr} only through its (typically negligible) contribution conditional on X_{rel}. We therefore characterize modality interference as the model retaining *excess* mutual information between Y and X_{irr} *after* conditioning on the task-relevant modality:

$$\mathcal{I}_{\text{interf}}(\theta) \;:=\; I_\theta(Y;\, X_{\mathrm{irr}} \,|\, X_{\mathrm{rel}})\,. \tag{10}$$

A model with strong cross-modality competency satisfies $\mathcal{I}_{\text{interf}}(\theta) \approx 0$, i.e., once X_{rel} is observed, X_{irr} provides no additional predictive signal.

Connection to the causal formulation. The causal effect δ_{cp} in Section 3 measures whether $\arg\max_y p_\theta(Y \mid X_{\mathrm{rel}}, X_{\mathrm{irr}})$ shifts under interventions on X_{irr}. This is a *decision-level* consequence of $\mathcal{I}_{\text{interf}}(\theta) > 0$: if Y retains nonzero conditional mutual information with X_{irr}, then there must exist interventions $\mathrm{do}(X_{\mathrm{irr}} = x'_{\mathrm{irr}})$ that change the predicted distribution. The two views are thus complementary: δ_{cp} is an observable, intervention-based diagnostic, while $\mathcal{I}_{\text{interf}}$ is the underlying information-theoretic quantity our training aims to suppress.

C.3 Conditional Information Bottleneck Objective

We pose the goal of mitigating modality interference as a constrained optimization problem in the spirit of the Information Bottleneck:

$$\min_\theta \; \underbrace{I_\theta(Y;\, X_{\mathrm{irr}} \mid X_{\mathrm{rel}})}_{\text{suppress irrelevant info}} \;-\; \beta \cdot \underbrace{I_\theta(Y;\, A \mid X_{\mathrm{rel}})}_{\text{preserve task signal}}, \tag{11}$$

where $\beta > 0$ trades off interference suppression against task fidelity. Equation (11) is intractable in general because both mutual information terms involve high-dimensional, implicit distributions induced by f_θ. We next show that the components of our training objective serve as a variational surrogate for Eq. (11).

C.4 Variational Surrogate via Perturbation and Consistency

Notation for perturbations. Let $\widetilde{X}_{\mathrm{irr}}$ denote a perturbed version of X_{irr}, drawn from a perturbation distribution $q(\widetilde{x}_{\mathrm{irr}} \mid x_{\mathrm{irr}})$ that includes both heuristic interventions and adversarial perturbations in the embedding space. We write $p_\theta^{\mathrm{clean}} := p_\theta(Y \mid X_{\mathrm{rel}}, X_{\mathrm{irr}})$ and $p_\theta^{\mathrm{pert}} := p_\theta(Y \mid X_{\mathrm{rel}}, \widetilde{X}_{\mathrm{irr}})$.

Upper bound on conditional mutual information. By the variational characterization of mutual information,

$$I_\theta(Y;\, X_{\mathrm{irr}} \mid X_{\mathrm{rel}}) \;=\; \mathbb{E}_{X_{\mathrm{rel}}, X_{\mathrm{irr}}}\big[\mathrm{KL}\big(p_\theta(Y \mid X_{\mathrm{rel}}, X_{\mathrm{irr}}) \,\big\|\, p_\theta(Y \mid X_{\mathrm{rel}})\big)\big]\,. \tag{12}$$

The marginal $p_\theta(Y \mid X_{\mathrm{rel}}) = \mathbb{E}_{X_{\mathrm{irr}}}[p_\theta(Y \mid X_{\mathrm{rel}}, X_{\mathrm{irr}})]$ is intractable to evaluate. Following standard variational practice, we replace it with a *perturbation-averaged predictive distribution*:

$$\bar{p}_\theta(Y \mid X_{\mathrm{rel}}) \;:=\; \mathbb{E}_{\widetilde{X}_{\mathrm{irr}} \sim q(\cdot \mid X_{\mathrm{irr}})}\Big[p_\theta(Y \mid X_{\mathrm{rel}}, \widetilde{X}_{\mathrm{irr}})\Big]\,. \tag{13}$$

Intuitively, if X_{irr} truly carries no task-relevant information, then perturbing it should not change the predictive distribution, and $\bar{p}_\theta$ should approximate $p_\theta(Y \mid X_{\mathrm{rel}})$ well.

By the convexity of KL divergence and Jensen's inequality,

$$\mathbb{E}_{X_{\mathrm{irr}}}\big[\mathrm{KL}\big(p_\theta^{\mathrm{clean}} \,\big\|\, \bar{p}_\theta(Y \mid X_{\mathrm{rel}})\big)\big] \;\le\; \mathbb{E}_{X_{\mathrm{irr}},\widetilde{X}_{\mathrm{irr}}}\big[\mathrm{KL}\big(p_\theta^{\mathrm{clean}} \,\big\|\, p_\theta^{\mathrm{pert}}\big)\big]\,. \tag{14}$$

Combining Eq. (12) and Eq. (14) (and using that $\bar{p}_\theta$ is a valid variational approximation to the true marginal under the assumption that q covers the support of perturbations that leave the task-relevant pathway intact), we obtain the upper bound

$$I_\theta(Y;\, X_{\mathrm{irr}} \mid X_{\mathrm{rel}}) \;\le\; \underbrace{\mathbb{E}_{X_{\mathrm{rel}},X_{\mathrm{irr}},\widetilde{X}_{\mathrm{irr}}}\big[\mathrm{KL}\big(p_\theta^{\mathrm{clean}} \,\big\|\, p_\theta^{\mathrm{pert}}\big)\big]}_{=:\,\mathcal{L}_{\mathrm{cons}}(\theta)} + C, \tag{15}$$

where C collects approximation terms that vanish when q leaves the task-relevant conditional invariant.[1] The right-hand side of Eq. (15) is precisely our consistency loss $\mathcal{L}_{\mathrm{consistency}}$ (Eq. (7)), instantiated with KL divergence (the JS instantiation yields an analogous symmetrized bound).

Adversarial perturbation as a tightness probe. The bound in Eq. (15) is loose when the perturbation distribution q fails to expose directions along which Y is most sensitive to X_{irr}. Our adversarial perturbation (Eq. (5)) explicitly seeks

$$\delta^\star \;=\; \arg\max_{\|\delta\|_\infty \le \epsilon} \mathrm{KL}\big(p_\theta(Y \mid X_{\mathrm{rel}}, X_{\mathrm{irr}}) \,\big\|\, p_\theta(Y \mid X_{\mathrm{rel}}, X_{\mathrm{irr}} + \delta \odot M)\big)\,, \tag{16}$$

where M is the modality-specific mask restricting δ to X_{irr}. Training under $\delta^\star$ therefore minimizes a *worst-case* version of the right-hand side of Eq. (15) within the ϵ-ball, tightening the variational bound on $I_\theta(Y;\, X_{\mathrm{irr}} \mid X_{\mathrm{rel}})$ where the model is most vulnerable to interference.

Task-fidelity term. The supervised loss $\mathcal{L}_{\mathrm{sft}}$ is the standard cross-entropy over ground-truth answer tokens. By the identity $H(A \mid Y, X_{\mathrm{rel}}) = H(A \mid X_{\mathrm{rel}}) - I(Y; A \mid X_{\mathrm{rel}})$, minimizing $\mathcal{L}_{\mathrm{sft}} \approx \mathbb{E}[-\log p_\theta(A \mid X_{\mathrm{rel}}, X_{\mathrm{irr}})]$ upper-bounds $H(A \mid Y, X_{\mathrm{rel}})$ and thus *maximizes* $I_\theta(Y; A \mid X_{\mathrm{rel}})$, recovering the second term of the conditional IB objective in Eq. (11).

C.5 Conclusion

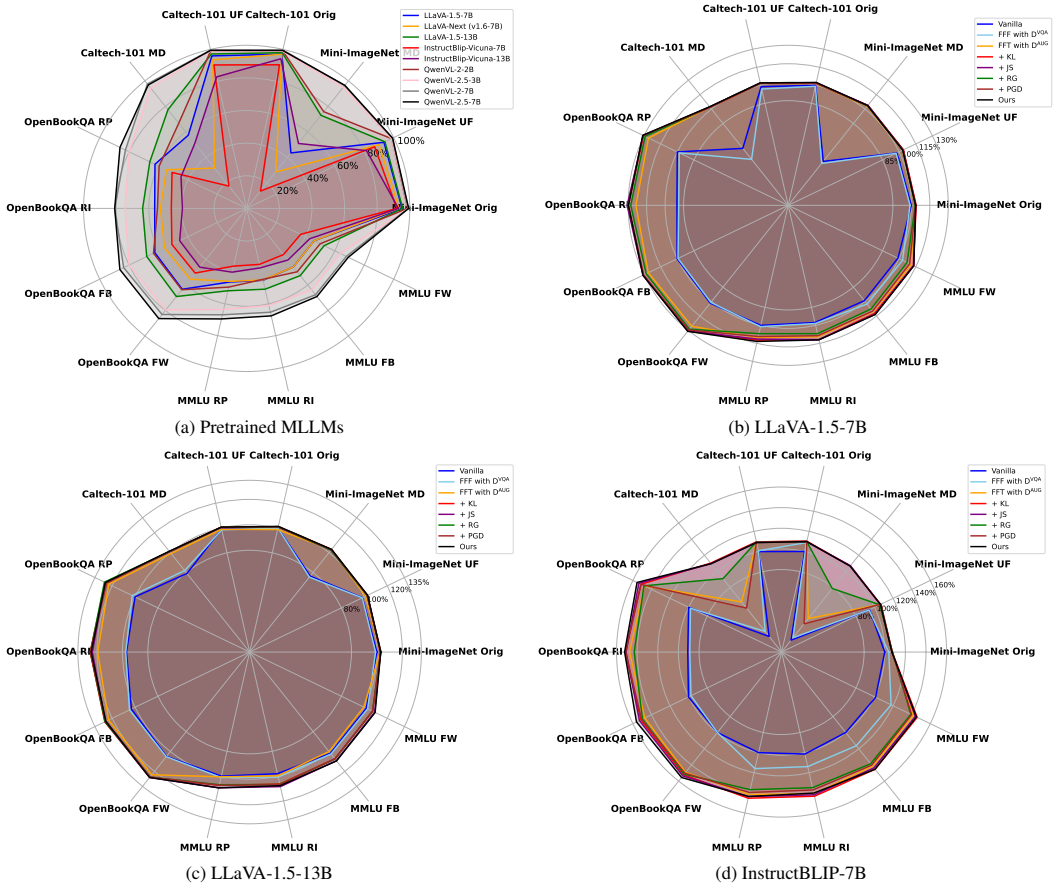
Combining the surrogate for the interference term (Eq. (15), tightened by adversarial search) with the cross-entropy lower bound on task-relevant mutual information, our total objective

$$\mathcal{L}_{\mathrm{total}} \;=\; \mathcal{L}_{\mathrm{sft}} \;+\; \mathcal{L}_{\mathrm{adv}} \;+\; \lambda_{\mathrm{cons}}\,\mathcal{L}_{\mathrm{consistency}} \tag{17}$$

constitutes a tractable variational surrogate for the conditional Information Bottleneck problem in Eq. (11), with λ_{cons} playing the role of the inverse Lagrange multiplier $1/\beta$. This interpretation does not replace the causal formulation; rather, it explains *why* the causal intervention δ_{cp} decreases under our training: minimizing $\mathcal{L}_{\mathrm{cons}} + \mathcal{L}_{\mathrm{adv}}$ provably upper-bounds the conditional mutual information through which X_{irr} can influence Y in the first place.

Limitations The bound in Eq. (15) relies on two assumptions: (i) the perturbation distribution q covers semantically meaningful variation in X_{irr} while preserving the task-relevant conditional, which we operationalize through diverse heuristic perturbations and modality-specific masking; and (ii) the gap term C is small, which is empirically supported by the observation that $\mathcal{L}_{\mathrm{cons}}$ correlates with reduced δ_{cp} across our experiments. We view this analysis as a principled *interpretation* of our framework rather than a tight theoretical characterization, and leave a fully variational bound (e.g., via explicit estimation of $\bar{p}_\theta$) to future work.

[1] Specifically, $C = \mathbb{E}[\mathrm{KL}(\bar{p}_\theta \,\|\, p_\theta(Y \mid X_{\mathrm{rel}}))]$, which is small when perturbations on X_{irr} approximately preserve the marginal predictive distribution—the very property our modality-specific masking is designed to enforce.



(a) Pretrained MLLMs

(b) LLaVA-1.5-7B

(c) LLaVA-1.5-13B

(d) InstructBLIP-7B

Figure 6: Task-wise robustness under perturbation. Each radar chart shows model accuracy (%) across Mini-ImageNet, Caltech-101 (image-heavy) and OpenBookQA, MMLU (text-heavy) under various perturbations. (a) uses raw accuracy of different pretrained MLLMs directly. (b–d) are normalized relative accuracy of each MLLMs. (We normalize each absolute accuracy into relative accuracy, which refers to absolute tested accuracy / accuracy of vanilla MLLMs in origin setting without perturbation.)

Table 14: Hyperparameter Settings Example

Category	Setting
Model and Training Strategy	
Base Model	LLaVA-1.5-7B
Finetune Type	Full
Adversarial Type	PGD-alike
Step size α	0.1
Epsilon-hall ϵ	0.001
Avdersarial Training Steps T	2
Consistency Regularization Type	JS
Loss Weight $\lambda_{\text{consistency}}$	0.01
Temperature τ	1
Optimization	
Epochs	1
Batch Size per GPU $\lvert\mathcal{B}\rvert$	8
Img/Text Ratio	0.25/0.25
Learning rate	2×10^{-5}

Table 15: Dataset Statistics

Dataset	Train	Test
Mini-ImageNet	4935	2000
Caltech-101	8124	1020
OpenBookQA	4957	1000
MMLU	7M	5469
LLaVA-Instruct	624610	-
TextCaps	109765	-
MMBench-EN	-	4377
ScienceQA-IMG	-	4114
SeedBench-IMG	-	14243

D Hyper-Parameter Setting and Training details

Please see Table 14 for more details.

D.1 Parameter Analysis with adversarial training iterations

To investigate the effect of adversarial strength on model performance, we vary the number of adversarial training iterations from 1 to 5 and evaluate the resulting VQA accuracy. As shown in Figure 7, both LLaVA-1.5-7B and LLaVA-1.5-13B models benefit from adversarial consistency training, with performance peaking at 2-step adversarial training (66.82% and 69.72%, respectively). Notably, excessive iterations (e.g., 4 or 5 steps) may lead to slight degradation, especially in larger models, likely due to over-perturbation and optimization difficulty.

These findings suggest that a moderate adversarial training setting (2 steps with ϵ=1e-3 and α=0.1 in LLaVA-1.5-7b, ϵ=1e-4 and α=0.1 in LLaVA-1.5-13b) offers an optimal balance between robustness and training stability, and that model size influences sensitivity to adversarial signal strength.

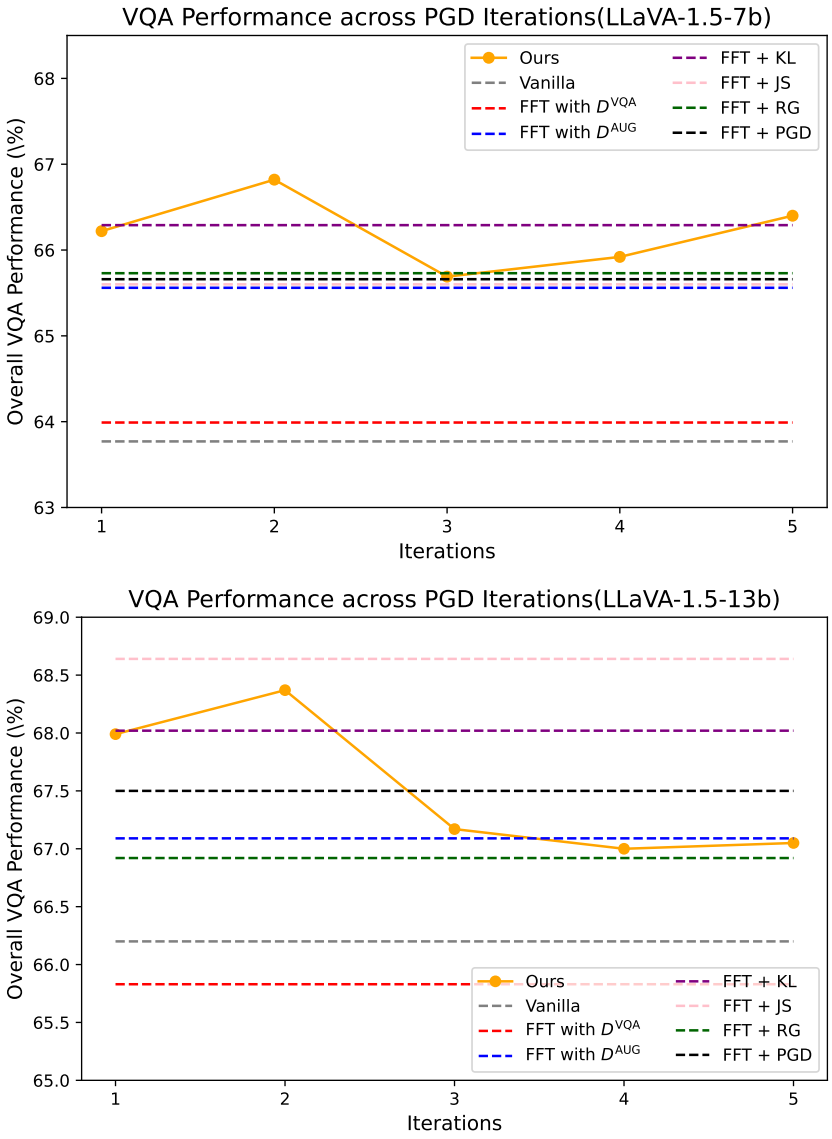


Figure 7: Comparison of VQA performance across adversarial training iterations for different model sizes.

E Experiments compute resources

All experiments were conducted on 8×A100 GPUs using DeepSpeed ZeRO-3 [57] with CPU offloading.

To quantify the computational overhead introduced by our adversarial training, we provide both theoretical FLOPs analysis and empirical wall-clock training time across model scales.

In the standard supervised fine-tuning (SFT) setting, the FLOPs per batch can be approximated as:

$$\text{FLOPs}_{\text{SFT}} \approx B_s \cdot (f_{\text{LLM}} + b_{\text{LLM}}), \tag{18}$$

where B_s is the batch size, f_{LLM} denotes the FLOPs of a forward pass through the LLM, and b_{LLM} the backward pass. In our adversarial training, each sample undergoes N adversarial training steps, each requiring an additional forward pass through the *frozen* LLM. Since gradients are computed only with respect to input embeddings via `torch.autograd.grad`, the overhead is minimal and thus ignored. After perturbation, clean and adversarial inputs are concatenated, resulting in a forward cost of $2B_s \cdot f_{\text{LLM}}$, followed by one backward pass. The total cost becomes:

$$\text{FLOPs}_{\text{ADV+SFT}} = B_s \cdot (N f_{\text{LLM}} + 2 f_{\text{LLM}} + b_{\text{LLM}}). \tag{19}$$

The relative overhead compared to vanilla SFT is:

$$\frac{N f + 2f + b}{f + b}. \tag{20}$$

Assuming $b_{\text{LLM}} \approx 2 f_{\text{LLM}}$, this simplifies to:

$$\frac{N + 4}{3}. \tag{21}$$

For our default $N = 1$, the theoretical FLOPs increase to approximately $1.66\times$ that of SFT.

We further report the actual training time across model scales, as shown in Table 16.

Table 16: Training time (in hours) across model scales. Δ denotes additional overhead per adversarial training iteration step.

Model	SFT	SFT + KL	SFT + ADV	Ours (ADV + KL)
Qwen2.5-VL-3B	1.5h	1.5–1.75h	3–3.5h (Δ = 0.5h)	3.5–4h (Δ = 0.5h)
LLaVA-1.5-7B	4h	4–4.5h	6–6.5h (Δ = 0.5h)	6–6.5h (Δ = 0.5h)
LLaVA-1.5-13B	10h	10–11h	13–14h (Δ = 1h)	13–14h (Δ = 1h)

Although the theoretical FLOPs suggest a $\sim$66% increase in cost when $N = 1$, the actual wall-clock time increase is much smaller. This is because our designed adversarial training leverages forward-only passes over frozen LLMs, avoiding costly backward and optimizer updates. As a result, the added runtime remains modest even on large models (e.g., only +2.5h for LLaVA-13B). Moreover, KL consistency training introduces negligible overhead compared to SFT.

F Limitations

Our analysis of modality interference is conducted from a coarse-grained perspective, primarily categorizing tasks into image-heavy and text-heavy types.

While our use of adversarial training represents a strong and generalizable perturbation strategy, it still operates within a defined input space (e.g., embedding-level noise bounded by L_∞ norms). Hence, adversarial perturbation should be viewed as a practical but partial solution rather than a comprehensive defense. Developing mechanisms that generalize across both semantic and modality perturbations remains an open and challenging direction.

G The Use of Large Language Models

We used large language models only to edit the manuscript for clarity, grammar, and academic style. No part of the research design, data analysis, or scientific content relied on language models, and the authors retain full responsibility for the paper's ideas and conclusions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's main contributions: (1) formalizing the Cross-Modality Competency Problem and identifying modality interference in MLLMs, (2) proposing a perturbation-based causal analysis to quantify modality reliance, and (3) introducing a fine-tuning strategy combining supervised augmentation, adversarial and heuristic perturbations, and consistency regularization.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the Section F

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present any formal theoretical results or proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides detailed descriptions of model architectures, datasets, training settings, and enabling reproduction of all main results in Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will include anonymized code and detailed instructions in the supplementary material to enable reproduction of all experiments. This includes data preprocessing steps, and scripts for training and evaluation.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 and Section D provide all training and evaluation details, including dataset splits, hyperparameters, optimization strategy, and implementation setup.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We ran each experiment five times and calculated the average.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section E reports GPU types, model scales, and training time for each setting. Total compute budget is estimated based on model size and data scale.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work complies with the NeurIPS Code of Ethics. All models and datasets used are publicly available, and we follow best practices for transparency, reproducibility, and safety.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section F discusses both positive impacts and potential risks such as model misuse via prompt manipulation.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not involve models or data with high risk for misuse. We only use publicly available MLLMs and datasets under standard academic licenses.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and models used in the paper are publicly released with proper citations and licenses.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: This work does not introduce new datasets or pretrained models. Only perturbation-augmented variants of existing data are used during training.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects are involved, so IRB approval is not required.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We leverage publicly available MLLMs (LLaVA, QwenVL, InstructBLIP) as core components of our diagnostic and fine-tuning framework, and their usage is detailed in Section 5.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.