Diagnosing and Mitigating Modality Interference in Multimodal Large Language Models

Rui Cai Bangzheng Li Xiaofei Wen Muhao Chen Zhe Zhao

UC DAVIS
UNIVERSITY OF CALIFORNIA

Paper · Code · Project page

MLLMs can't tell which modality to trust: an irrelevant caption or image can flip their answer, even at 100B+ scale.

1 Motivation

A **misleading caption** flips an image-classification answer; a **blank image** changes a pure-text QA answer.

Cross-Modality Competency

The ability to **fairly evaluate and integrate** all modalities: rely on the *task-relevant* one, ignore the irrelevant one.

prior work: separate diagnoses

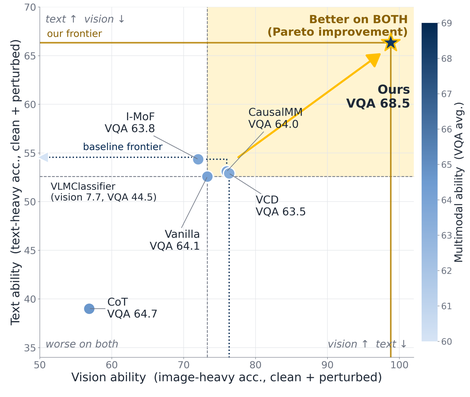
Catastrophic forgetting | Knowledge conflict | Spurious correlations

one root cause: **deficient Cross-Modality Competency**

Modality Interference
irrelevant modality changes the answer

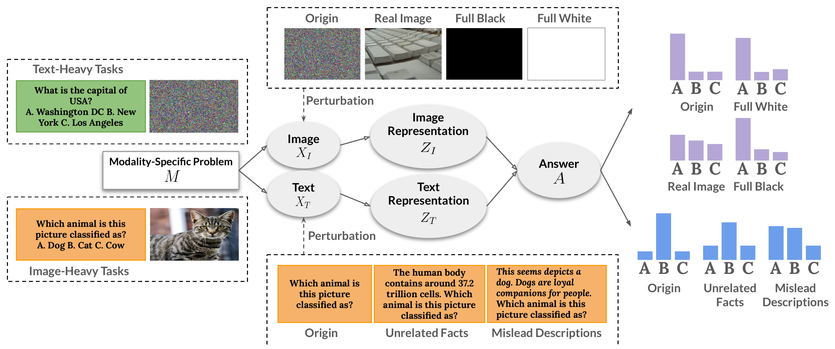
Goal: Pareto Optimality

Improve vision, text **and** multimodal ability together: **push the frontier**, don't trade along it.



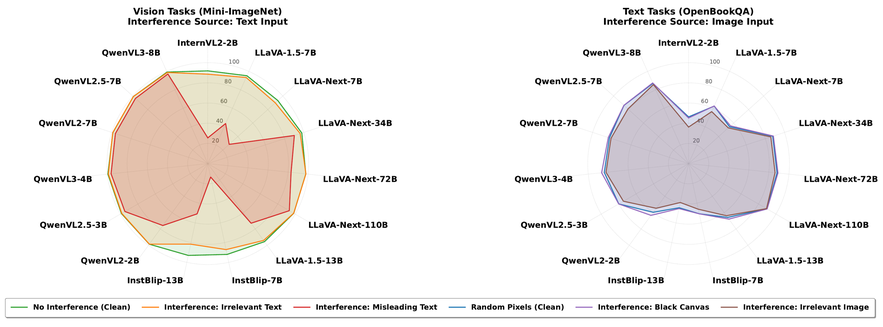
LLaVA-1.5-7B (Table 1). Color = VQA average; quadrants centred on vanilla.

2 Diagnosing Modality Interference



Intervene **only on the task-irrelevant modality**; count how often the answer flips (δ_{cp}).

Result: interference is pervasive across 15 MLLMs (2B–110B)



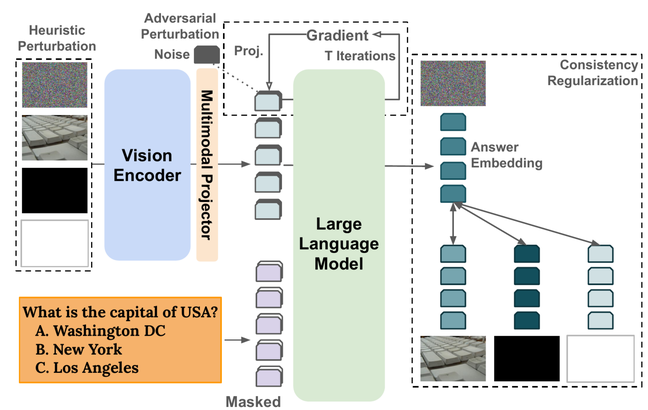
92.0 → 13.6
InstructBLIP-7B, Mini-ImageNet + misleading text

95.3 → 43.5
LLaVA-1.5-7B, Mini-ImageNet + misleading text

85.9 → 77.5
Qwen2.5-VL-7B, OpenBookQA + unrelated image

Scale hides it, but does not remove it: a black image leaves Qwen3-VL-8B's accuracy flat (87.0 → 87.5), yet **5.6%** of answers flip.

3 Method: Perturbation-Aware Fine-Tuning



1. **Heuristic perturbation**
noisy copies of each sample on the irrelevant modality

2. **Masked adversarial perturbation**
PGD on irrelevant-modality token embeddings only

3. **Consistency regularization**
same output on clean and perturbed inputs (KL / JS)

4 vs. Baselines

Prompting, decoding and architecture fixes leave interference.

	Image (pert.)		Text (pert.)		VQA		
Method	**MiniIN**	**Caltech**	**OBQA**	**MMLU**	**SciQA**	**MMB**	**SEED**
LLaVA-1.5-7B							
Vanilla	43.5	57.4	56.4	45.2	64.5	64.3	63.4
+ CoT	28.9	36.0	38.9	38.7	64.7	65.2	64.1
+ VLMCls.	0.0	0.0	61.2	47.8	61.8	35.8	36.0
+ I-MoF	41.2	58.1	63.8	40.0	62.1	65.2	64.2
+ CausalMM	47.2	64.5	56.6	47.0	63.9	64.3	63.7
+ VCD	50.9	63.3	56.4	46.6	63.0	63.8	63.8
Ours	**98.4**	**98.9**	**81.0**	**51.0**	**67.8**	**73.1**	**64.6**
LLaVA-1.5-13B							
Vanilla	73.0	77.4	63.8	50.8	65.8	72.1	64.5
+ CoT	62.8	67.8	53.0	45.5	65.6	70.2	64.9
+ VLMCls.	0.0	0.0	65.3	51.3	66.7	56.2	55.3
+ I-MoF	70.1	80.9	64.5	42.1	**66.8**	73.0	66.6
+ CausalMM	75.1	83.5	58.1	50.3	64.1	72.0	64.1
+ VCD	75.7	82.4	63.4	50.6	66.3	70.9	63.3
Ours	**98.4**	**98.6**	**82.1**	**55.8**	65.3	**73.7**	**68.4**

Perturbed = worst-case perturbation. Bold = best per backbone.

5 Generalization

Unseen noise (OCR, UI screenshots) and unseen datasets.

Model		Caltech +OCR	OBQA +Screen	Food-101	Food +OCR	ARC
LLaVA-1.5-7B	Vanilla	92.8	50.2	90.2	83.6	53.4
	Ours	**99.0**	**81.2**	**92.1**	**89.5**	**62.8**
InstructBLIP-7B	Vanilla	83.6	40.0	67.9	48.0	33.7
	Ours	**99.0**	**77.8**	**95.6**	**88.1**	**57.6**
Qwen2.5-VL-7B	Vanilla	99.2	69.3	97.1	96.7	**85.6**
	Ours	**99.5**	**90.7**	**97.5**	**97.4**	84.1

17.5 → 99.0
InstructBLIP-7B Caltech, perturbed

40.0 → 77.8
InstructBLIP-7B OBQA + screenshot