MoDA

Modulation Adapter for Fine-Grained Visual Understanding in Instructional MLLMs

Wayner Barrios[1], Andrés Villa[2], Juan C. Leon Alcazar[2], SouYoung Jin[1 †], Bernard Ghanem[2 †]

[1] Department of Computer Science, Dartmouth College, [2]King Abdullah University of Science and Technology

[†] *Equal senior authorship.*

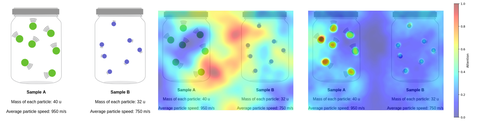
Motivation

- **MLLMs struggle with fine-grained visual understanding**, causing hallucinations on queries that need precise localization.
- **Cause:** CLIP patches suffer **semantic entanglement**, so feature dimensions blend unrelated concepts.
- Prior fixes add encoders/retraining or **ignore the instruction**.
- **MoDA**: instruction-guided channel-wise modulation with no architecture changes and <1% FLOPs.

Contributions

- **MoDA** is a lightweight adapter that dynamically modulates visual features via language cues.
- **MoDA** delivers consistent gains across 12 benchmarks and three MLLM families (LLaVA-1.5, LLaVA-MoRE, Qwen3-VL). **Largest gains on fine-grained visual tasks**: +12.0 MMVP, +4.9 ScienceQA (Qwen3-VL).

Qualitative Results



(a) Input Image (b) Baseline (LLaVA-MoRE) **(c) Ours (LLaVA-MoRE + MoDA)**

Did the snake stick out its tongue and touch its skin? (a) Yes (b) No

LM ✓ Yes, the snake has its tongue sticking out and touching its skin.

MoDA ✓ Yes, the snake has its tongue sticking out and touching its skin.

Did the snake stick out its tongue and touch its skin? (a) Yes (b) No

LM ✗ Yes, the snake has its tongue sticking out and touching its skin.

MoDA ✓ No

Are there any words displayed on the vehicle's lightbar? (a) Yes (b) No

LM ✓ Yes, there are words displayed on the vehicle's lightbar.

MoDA ✓ Yes

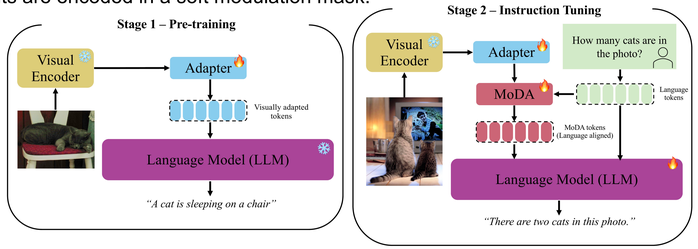
Are there any words displayed on the vehicle's lightbar? (a) Yes (b) No

LM ✗ Yes, there are words displayed on the vehicle's lightbar.

MoDA ✓ No

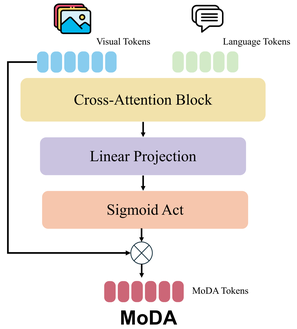
MoDA Model

MoDA (Modulation Adapter) is a lightweight module designed to post-process visual embeddings from an MLLM's adapter. MoDA leverages the alignment of visual and language embedding spaces and selects the most relevant visual features based on the input language query. Our module assigns individual weights to these visual features through cross-attention with the language embedding, these weights are encoded in a soft modulation mask.



(a) Stage 1 - Pre-Training

(b) Stage 2 - Instruction Tuning



Performance on VQA benchmarks

Method	LLM	GQA	ScienceQA	MMBench	RealWorldQA	ChartQA
BLIP-2 (Li et al., 2023a)	FLAN-T5	41.0	61.0	-	22.4	-
InstructBLIP (Dai et al., 2023)	Vicuna-7B	42.9	60.5	36.0	1.0	0.2
Qwen-VL-Chat (Bai et al., 2023)	Qwen-7B	57.5	68.2	60.6	-	-
LLaVA (Liu et al., 2023a)	Vicuna-7B	-	38.5	34.1	11.0	-
LLaVA-1.5 (Liu et al., 2024)	Vicuna-13B	63.3	71.6	67.7	45.8	17.1
ShareGPT-4V (Chen et al., 2024a)	Vicuna-7B	63.3	68.4	68.8	52.0	16.8
LLaVA-1.5 (Liu et al., 2024)	Vicuna-7B	62.4	69.0	64.3	44.3	**17.0**
LLaVA-1.5 + **MoDA (ours)**	Vicuna-7B	**62.5**	**71.0**	**64.8**	**53.4**	13.2
LLaVA-MoRE OpenAI CLIP (Cocchi et al., 2025)	LLaMA 3.1-8B	63.6	76.3	**72.3**	57.1	15.5
LLaVA-MoRE OpenAI CLIP + **MoDA (ours)**	LLaMA 3.1-8B	**64.4**	**77.8**	72.0	**58.0**	**15.6**
LLaVA-MoRE SigLIP-S2 (Cocchi et al., 2025)	LLaMA 3.1-8B	64.9	77.1	71.8	57.2	17.3
LLaVA-MoRE SigLIP-S2 + **MoDA (ours)**	LLaMA 3.1-8B	**65.4**	**81.9**	**72.4**	**58.2**	**18.1**
Qwen3-VL-2B-Instruct (Qwen Team, 2025)	Qwen3-2B	59.4	79.3	86.5	64.7	**80.0**
Qwen3-VL-2B-Instruct + **MoDA (ours)**	Qwen3-2B	**63.2**	**84.2**	**87.4**	**68.8**	79.0

Reducing Hallucination: POPE, MMVP

MMVP evaluates hallucination using 150 image pairs and 300 corresponding questions. The image pairs have a highly similar CLIP embeddings.

Method	LLM	POPE	MMVP^{all}
BLIP-2 (Li et al., 2023a)	FLAN-T5	-	-
InstructBLIP (Dai et al., 2023)	Vicuna-7B	85.0	16.9
LLaVA (Liu et al., 2023a)	Vicuna-7B	-	6.6
LLaVA-1.5 (Liu et al., 2024)	Vicuna-13B	85.9	24.7
LLaVA-1.5 (Liu et al., 2024)	Vicuna-7B	85.6	24.0
LLaVA-1.5 + **MoDA**	Vicuna-7B	**87.1**	**36.0**
LLaVA-MoRE OpenAI CLIP (Cocchi et al., 2025)	LLaMA-8B	85.1	27.3
LLaVA-MoRE OpenAI CLIP + **MoDA**	LLaMA-8B	**86.3**	**28.7**
LLaVA-MoRE SigLIP-S2 (Cocchi et al., 2025)	LLaMA-8B	86.0	39.3
LLaVA-MoRE SigLIP-S2 + **MoDA**	LLaMA-8B	**87.7**	**42.7**
Qwen3-VL-2B-Instruct (Qwen Team, 2025)	Qwen3-2B	89.4	-
Qwen3-VL-2B-Instruct + **MoDA**	Qwen3-2B	**89.9**	-

Ablation Study of MoDA Components.

MoDA Type	Supp. Loss	LLM	Vision Encoder	POPE	GQA	SQA	MMVP	Avg.
Baseline Models (No MoDA)								
-	-	Vicuna-7B	CLIP ViT-L/14@336	85.6	62.4	69.0	24.0	60.3
-	-	LLaMA 3.1-8B	CLIP ViT-L/14@336	85.1	63.6	76.3	27.3	63.1
-	-	LLaMA 3.1-8B	SigLIP-S2	86.0	64.9	77.1	39.3	66.8
CLIP ViT-L/14@336 Ablations								
Linear (MLP)	L_1	LLaMA 3.1-8B	CLIP ViT-L/14@336	87.2	64.3	76.7	28.7	64.2
Linear (MLP)	None	LLaMA 3.1-8B	CLIP ViT-L/14@336	86.6	64.4	77.8	28.1	64.2
Cross-Attention	L_1	LLaMA 3.1-8B	CLIP ViT-L/14@336	87.6	64.2	76.8	20.2	62.2
Self-Attention	None	LLaMA 3.1-8B	CLIP ViT-L/14@336	86.5	64.2	77.3	27.9	64.0
Cross-Attention	None	LLaMA 3.1-8B	CLIP ViT-L/14@336	86.3	64.4	77.8	28.7	64.3
LLM Backbone Comparison								
Cross-Attention	None	Vicuna-7B	CLIP ViT-L/14@336	87.1	62.5	71.0	36.0	64.2
SigLIP-S2 Ablations								
Linear (MLP)	L_1	LLaMA 3.1-8B	SigLIP-S2	85.8	65.2	77.9	39.6	67.1
Linear (MLP)	None	LLaMA 3.1-8B	SigLIP-S2	86.6	64.8	77.8	40.0	67.3
Cross-Attention	L_1	LLaMA 3.1-8B	SigLIP-S2	87.0	65.1	79.2	41.1	68.1
Self-Attention	None	LLaMA 3.1-8B	SigLIP-S2	**87.9**	64.9	79.9	39.5	68.0
Cross-Attention	None	LLaMA 3.1-8B	SigLIP-S2	87.7	**65.4**	**81.9**	**42.7**	**69.4**

Acknowledgments

This work was supported by startup funds provided by Dartmouth College, and by funding from the KAUST Center of Excellence for Generative AI under award number 5940. Computational resources were provided by IBEX, managed by the Supercomputing Core Laboratory at KAUST.

Paper Code