# MoDA: Modulation Adapter for Fine-Grained Visual Understanding in Instructional MLLMs

**Anonymous authors**
Paper under double-blind review

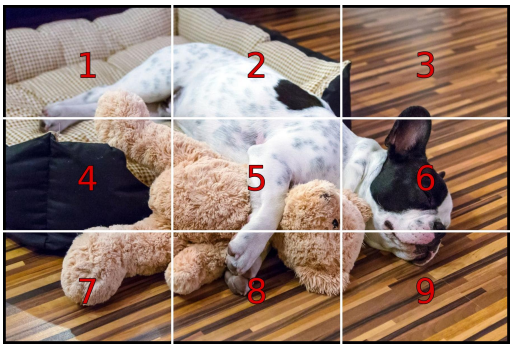
## Abstract

Multimodal Large Language Models (MLLMs) have achieved remarkable success in instruction-following tasks by integrating pretrained visual encoders with large language models (LLMs). However, existing approaches often struggle with fine-grained visual grounding due to semantic entanglement in visual patch representations, where individual patches blend multiple distinct visual elements, making it difficult for models to focus on instruction-relevant details. To address this challenge, we propose MoDA (Modulation Adapter), a lightweight module that enhances visual grounding through instruction-guided channel-wise modulation. Following the standard LLaVA training protocol, MoDA operates in the second stage by applying cross-attention between language instructions and pre-aligned visual features, generating dynamic modulation masks that emphasize semantically relevant embedding dimensions while de-emphasizing irrelevant information. This targeted refinement enables more precise visual-language alignment without architectural modifications or additional supervision. We conduct comprehensive evaluation across 13 diverse benchmarks spanning visual question answering, vision-centric reasoning, and hallucination detection. MoDA demonstrates substantial improvements, achieving notable gains of +12.0 points on MMVP hallucination detection and +4.8 points on ScienceQA reasoning, while consistently outperforming baselines on 12 out of 13 benchmarks with minimal computational overhead ($< 1\%$ FLOPs). Our results establish MoDA as an effective, general-purpose enhancement for improving fine-grained visual grounding in instruction-tuned MLLMs.

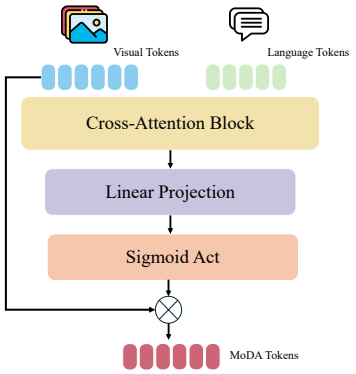
## 1 Introduction

The rapid progress of Large Language Models (LLMs) has led to impressive zero-shot performance across a broad spectrum of natural language processing benchmarks (Wang et al., 2024; Chung et al., 2024; Liang et al., 2023; Llama Team, AI @ Meta, 2024; Yang et al., 2024; Team, 2025). The success of instruction-tuned LLMs has driven computer vision research in a similar direction, ultimately leading to the development of Multimodal Large Language Models (MLLMs). MLLMs integrate pretrained visual encoders with large language models via lightweight adapter modules, enabling efficient cross-modal alignment and strong performance across diverse multimodal tasks, including Visual Question Answering (VQA), Image Captioning, Image Reasoning, and Image Classification.

Despite their success, state-of-the-art MLLMs frequently struggle with fine-grained visual understanding, particularly when answering queries that require precise localization and detailed reasoning about specific visual elements. This limitation manifests as hallucinations, where model outputs contradict actual image semantics, undermining reliability in real-world applications. Prior analyses have identified the CLIP-based visual encoder as a key bottleneck: its patch-based representations often fail to capture localized details due to semantic entanglement within individual patches (Villa et al., 2024; Tong et al., 2024b; Kar et al., 2024). While some works incorporate multiple specialized visual encoders (Tong et al., 2024b; Kar et al., 2024) or fine-tune CLIP for better local structure preservation (Villa et al., 2025), these approaches often introduce substantial computational overhead or require large-scale retraining.




(a) Image patches for ViT input representation




(b) Modulation Adapter (MoDA) Architecture

Figure 1: **ViT patch representation and our proposed Modulation Adapter (MoDA). (a)** ViT splits images into fixed-size patches, each projected into high-dimensional embeddings. This partitioning blends semantically distinct elements (e.g., dog, toy, floor within a single patch), creating entangled representations. **(b)** MoDA is a lightweight module that modulates visual embeddings via cross-attention using language tokens as guidance, enabling selective attention without architectural modifications or additional supervision.

We illustrate this semantic entanglement problem through a practical example. Figure 1a shows a $3\times3$ grid over a sleeping French bulldog with a plush toy, simulating CLIP's visual tokenization with enlarged patches for visualization. Crucially, none of the patches contain uniform visual elements. Patch 5 blends the dog's torso, stuffed toy, and cushioned bed; patch 6 mixes the dog's head, ear, and hardwood floor. This forces the visual encoder to combine distinct shapes, textures, and semantic concepts into single embeddings, where individual feature dimensions encode multiple unrelated meanings (Oquab et al., 2024; Ma et al., 2022; Zhou et al., 2024; Shi et al., 2024). Consequently, when processing language queries like *"What color is the dog's ear?"* or *"Is the toy lying on the bed or the floor?"*, the model must disentangle mixed visual representations to provide reliable answers, often failing to focus on instruction-relevant details.

Existing approaches to address this challenge fall into several categories. Some works apply attention masking techniques adapted from NLP (Fan et al., 2021; Tang et al., 2021; Lin & Joe, 2023; Rende et al., 2024), but these typically operate on token-level sparsity rather than channel-wise feature refinement. Others employ layer-wise adaptive masking (Barrios & Jin, 2024), which introduces substantial overhead when applied to deep models. Most critically, these approaches lack instruction-guided conditioning, missing the opportunity to dynamically adapt visual attention based on specific language queries. This leads to our central question: *How can we enable MLLMs to dynamically focus on instruction-relevant visual details for better visual understanding without architectural modifications or computational overhead?*

We address this challenge through the *Modulation Adapter* (**MoDA**), a lightweight module that performs instruction-guided channel-wise modulation of pre-aligned visual features. Unlike prior masking approaches (Barrios & Jin, 2024; Lin et al., 2022) that operate on attention weights or token-level sparsity, MoDA applies targeted modulation to visual embedding dimensions, emphasizing channels relevant to the current language instruction while de-emphasizing irrelevant information. Our approach employs cross-attention between language instructions and visual features to generate dynamic modulation masks, enabling precise visual-language alignment without modifying the underlying MLLM architecture. Crucially, MoDA's effectiveness scales with visual encoder quality: while providing modest improvements with standard CLIP encoders, it achieves substantial gains when paired with richer representations like SigLIP-S2, demonstrating that instruction-guided modulation becomes increasingly valuable for fine-grained visual understanding. MoDA integrates seamlessly into existing two-stage instruction-tuning pipelines, requires no additional supervision or training data, and introduces minimal computational overhead ($< 1\%$ FLOPs, $3.7\%$ parameters).

We validate MoDA across 13 diverse benchmarks spanning visual question answering, vision-centric reasoning, and hallucination detection using strong MLLM baselines (LLaVA-1.5 (Liu et al., 2024) and LLaVA-MoRE (Cocchi et al., 2025)). MoDA achieves substantial improvements in fine-

grained visual understanding, with **+12.0 points** on MMVP hallucination detection and **+4.8 points** on ScienceQA reasoning, outperforming baselines on **12 out of 13 benchmarks**. Ablation studies confirm these gains stem from architectural design rather than parameter scaling, with strongest improvements on fine-grained visual tasks. Our main contributions are: **(i)** identifying semantic entanglement in visual patch representations and proposing MoDA, a novel instruction-guided channel-wise modulation approach that addresses this limitation; **(ii)** demonstrating substantial performance improvements with minimal computational overhead, adding only $< 1\%$ FLOPs while achieving consistent gains across diverse benchmarks; and **(iii)** comprehensive evaluation showing MoDA's effectiveness stems from architectural innovation rather than capacity increases.

## 2 Related Work

**Multimodal Instruction Tuning.** Instruction-tuning has become the standard approach for enhancing MLLMs by incorporating task-specific natural language commands that improve generalization across vision-language tasks. The typical pipeline involves two stages: first, cross-modal alignment projects visual features from encoders like CLIP (Liu et al., 2023a; 2024; Cocchi et al., 2025; Chen et al., 2024a) or Q-Former (Li et al., 2023a; Dai et al., 2023) into the language embedding space; second, instruction-following fine-tuning enhances task generalization. Our approach builds upon the second stage, assuming well-aligned multimodal representations and focusing on instruction-conditioned refinement of visual features.

**Cross-Modal Attention and Feature Aggregation.** Modern MLLMs increasingly leverage cross-attention mechanisms for multimodal integration. InstructBLIP (Dai et al., 2023) pioneered injecting language queries directly into Q-Former architecture for selective visual attention, while Cambrian-1 (Tong et al., 2024a) employs cross-attention at the token level for multimodal reasoning. Other approaches explore multiple visual encoders with cross-attention fusion (Kar et al., 2024) or learnable query tokens for task-relevant information extraction. However, these methods primarily operate on discrete token interactions. MoDA differs by introducing channel-wise modulation through cross-attention, where language instructions guide the re-weighting of continuous feature dimensions rather than discrete tokens, enabling fine-grained semantic control while preserving the spatial structure of visual representations.

**Attention Masking and Multimodal Efficiency.** Attention masking strategies in multimodal models can be categorized into three main paradigms. Token-level sparsity methods like SwinBERT (Lin et al., 2022) generate fixed sparse masks at input, trading adaptability for efficiency. Layer-wise adaptive approaches such as LAM (Barrios & Jin, 2024) recompute learnable masks at each transformer layer, enabling dynamic attention but introducing computational overhead that scales problematically with network depth. Visual-only mechanisms like MST (Li et al., 2021) perform attention-guided masking within the vision encoder without language interaction. MoDA introduces a distinct fourth paradigm through single-pass channel-wise modulation that operates on continuous feature dimensions rather than discrete tokens, performs modulation only once after the adapter stage to avoid scaling issues, and explicitly incorporates language guidance for instruction-conditioned refinement.

**Adapter Architectures.** Adapter modules serve as crucial interfaces between visual encoders and language models in MLLMs. While LLaVA-family models (Liu et al., 2023a; Cocchi et al., 2025; Chen et al., 2024a) employ lightweight adapters for efficient CLIP-to-language mapping, recent innovations include attention pooling and multi-scale feature aggregation. However, these approaches primarily focus on initial cross-modal alignment rather than dynamic, instruction-conditioned refinement. MoDA complements existing adapter architectures by operating as a post-processing module that refines already-aligned features based on specific language instructions, maintaining compatibility with standard MLLM designs while providing targeted improvements in fine-grained visual grounding.

**Visual Feature Refinement Across the Pipeline.** Recent work has explored visual feature refinement at different stages of the MLLM pipeline. At the encoder level, EAGLE (Villa et al., 2025) fine-tunes CLIP to better preserve local structure, requiring additional pre-training. Instruction-Guided Fusion (Li, 2025) addresses layer selection by dynamically weighting features from different encoder depths based on task requirements. At the decoder level, MoReS (Bi et al., 2024) applies linear transformations at each LLM layer to address modality imbalance where text dominates visual representations. AdaLink (Wang et al., 2023) introduces input-centric parameter-efficient fine-

tuning through non-intrusive adaptation mechanisms. These methods operate at distinct pipeline stages: encoder pre-training, layer selection, or per-layer LLM transformations. In contrast, MoDA operates at the adapter-to-LLM interface, performing channel-wise modulation on already-aligned features before they enter the language model. This positioning makes MoDA potentially complementary to the above approaches, as improved encoder features or layer selection could provide higher-quality inputs for MoDA's channel-wise refinement, while MoDA's instruction-conditioned modulation could enhance the features before downstream processing by methods operating within the LLM.

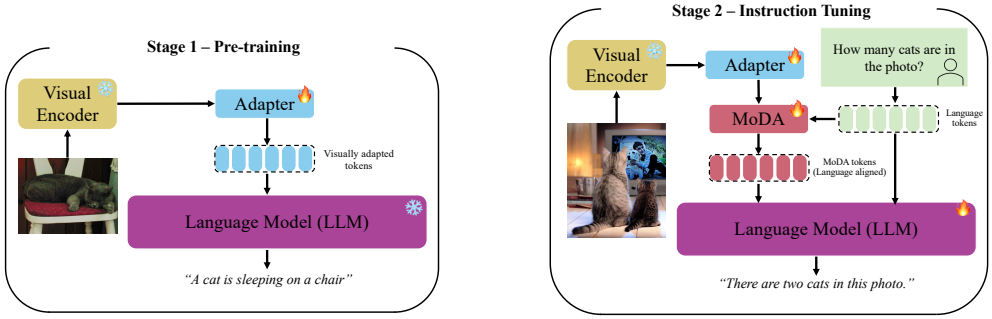
## 3 Visual Feature Modulation

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. This mask promotes relevant visual embedding dimensions while de-emphasizing less relevant ones. The resulting re-weighted features are then passed to the LLM for decoding.

Within a MLLM, the MoDA component is integrated after the pre-trained adapter. Given a pre-aligned visual feature map $V_{\text{aligned}}$, our objective is to learn a function $F(\cdot)$ that estimates a modulation operator based on the current text query $T$. This operator is then applied element-wise across the embedding dimensions of the visual features, as follows:

$$\widetilde{V}_{\text{aligned}} = V_{\text{aligned}} \odot F(T, V_{\text{aligned}}) \tag{1}$$

Where $\odot$ denotes the Hadamard product along the embedding dimension. The function $F(T, V_{\text{aligned}})$ is dependent on the text prompt, therefore, it modulates the attention of the MLLM towards the more informative embeddings according to the current text prompt. As a consequence, the re-weighted feature map $\widetilde{V}_{\text{aligned}}$ provides refined visual cues, which improve the MLLM's ability to resolve the complex natural language instructions in modern MLLM benchmarks.




(a) Stage 1 - Pre-Training (b) Stage 2 - Instruction Tuning

Figure 2: **Training Framework.** MoDA follows a two-stage process: **(1) Pre-training** the adapter for visual–language alignment, and **(2) Instruction Tuning** where the adapter and MoDA are fine-tuned with a pretrained LLM. MoDA refines adapter outputs by emphasizing language-relevant visual features.

### 3.1 Modulation Adapter (MoDA) Design

Let $V_{\text{aligned}} \in \mathbb{R}^{B\times N\times E}$ denote the language aligned visual features obtained from the adapter the module of the MLLM, where $B$ is the batch size, $N$ is the number of image tokens, and $E$ is the embedding dimension. Let $T \in \mathbb{R}^{B\times M\times E}$ represent the language token embeddings, where $M$ is

the number of text tokens. The $T$ embeddings are obtained directly from the initial layers of the LLM component. MoDA learns a modulation function $F(\cdot) \in [0,1]^E$ conditioned on the multi-modal feature embedding $\{V_{\text{aligned}}, T\}$, followed by a linear projection and sigmoid activation. The re-weighted visual features $\widetilde{V}_{\text{aligned}}$ are computed as:

$$\widetilde{V}_{\text{aligned}} = V_{\text{aligned}} \odot \sigma\left(W \cdot F(T, V_{\text{aligned}})\right) \tag{2}$$

The modulation function $F(\cdot)$ is implemented using a stack of Transformer Layers that takes the language-aligned visual features $V_{\text{aligned}}$ as the target sequence and the language token embeddings $T$ as the memory input. The matrix $W \in \mathbb{R}^{E \times E}$ is a learnable linear projection, and $\sigma(\cdot)$ is the sigmoid activation function applied element-wise to constrain the mask values in the range $[0, 1]$. In practice, the output of $\sigma(W \cdot F(T, V_{\text{aligned}}))$ can be interpreted as a channel-wise mask that independent re-weights each feature channel in the visual embedding.

The MoDA module consists of multiple cross-attention Transformer layers, each composed of three main components: (i) a multi-head cross-attention mechanism that allows each visual token to attend to relevant parts of the language input, (ii) a feed-forward network that refines the representation at each layer, and (iii) residual connections and layer normalization to facilitate training stability and convergence. After passing through this stack, the output is projected and passed through the sigmoid non linearity to generate the final modulation mask $\mathcal{M}$. This mask is applied following equation 1 to obtain the refined visual representation $\widetilde{V}_{\text{aligned}}$.

### 3.2 MoDA MLLM Architecture and Training Details

MLLMs incorporating with MoDA adopt the architecture and two-stage training protocol introduced in LLaVA Liu et al. (2023a), which ensembles a vision encoder with a large language model (LLM). As illustrated in Figure 2, our enhanced MLLM retains the three fundamental components of Liu et al. (2023a): a vision encoder, an adapter module for visual-language alignment, and a pretrained LLM. However, MoDA (Modulation Adapter) is introduced as a novel component that operates as an interface between the pre-trained vision-language adapter and the LLM. Following this integration, the vision encoder extracts patch-level visual features from the input image, which are then projected into the language embedding space by the standard adapter module. MoDA then takes these aligned visual features, estimates channel-wise modulation weights, and passes the modulated features to the LLM for language decoding.

Following the standard practice in LLaVA models, the enhanced visual embeddings are then used as prefix tokens for the LLM. Then, LLM mixes the modulated visual tokens with the input query tokens, and autoregressively generates a natural language response.

**Training Procedure.** The training of MoDA follows the two-stage approach of Liu et al. (2023a). In the first stage, only the original visual adapter is trained following the LLaVA protocol Liu et al. (2023a; 2024). The vision encoder and the LLM remain frozen during this phase, and the training is supervised using an autoregressive language modeling objective. The LLM is prompted with language-aligned image features (via the adapter) and a language instruction, and it learns to predict the target output sequence using standard cross-entropy loss over the predicted tokens.

In the second stage, we introduce the MoDA module to enhance the model's grounding capabilities. MoDA is initialized using Xavier initialization, while the visual adapter retains the weights learned on the initial stage. During this phase, we finetune both MoDA and the LLM jointly, enabling the model to better attend to semantically relevant visual cues through MoDA while improving its overall conversational ability.

The learning objective across both stages reamains the same: given a sequence of input tokens and visual embeddings, the model is trained to minimize the autoregressive cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, \widetilde{V}_{\text{aligned}}, T) \tag{3}$$

where $y_t$ is the ground-truth token at time step $t$, $y_{<t}$ denotes the previously generated tokens, $\widetilde{V}_{\text{aligned}}$ are the modulated visual features produced by MoDA, and $T$ represents the tokenized instruction.

# 4 EXPERIMENTS

Our experimental evaluation strategically targets the semantic entanglement problem identified in Figure 1 through 13 benchmarks spanning three categories: hallucination detection where models must distinguish visual evidence from learned priors, complex reasoning requiring precise visual-language coordination, and fine-grained visual analysis demanding detailed instruction-following capabilities.

**Experimental Setup.** We evaluate MoDA across 13 benchmarks spanning visual question answering (GQA, ScienceQA, MMBench variants, RealWorldQA, ChartQA), vision-centric tasks (LLaVA-Wild, MMVet, MMStar, V*Bench, CV-Bench), and hallucination detection (POPE, MMVP). These benchmarks require strong language capabilities for instruction following and precise visual processing. Our model follows the standard LLaVA architecture with MoDA integrated as a lightweight cross-attention module between the adapter and language model. We adopt the two-stage training protocol of LLaVA-1.5, using the same hyperparameters and training data to ensure fair comparison. More details in Appendix section (Section A.1).

Table 1: **Performance on Visual Question Answering benchmarks.** We evaluate on GQA, ScienceQA, MMBench (En/Cn), RealWorldQA, and ChartQA. **Bold underlined** values indicate highest scores per benchmark. **Bold** values show best performance within each baseline comparison. Gray text indicates models trained on different larger data distributions. All metrics are percentages; higher is better.

| Method | LLM | GQA | ScienceQA | MMBench-En | MMBench-Cn | 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 | 23.7 | 1.0 | 0.2 |
| Qwen-VL-Chat (Bai et al., 2023) | Qwen-7B | 57.5 | 68.2 | 60.6 | 56.7 | - | - |
| LLaVA (Liu et al., 2023a) | Vicuna-7B | - | 38.5 | 34.1 | 14.1 | 11.0 | - |
| LLaVA-1.5 (Liu et al., 2024) | Vicuna-13B | 63.3 | 71.6 | 67.7 | 63.6 | 45.8 | 17.1 |
| ShareGPT-4V (Chen et al., 2024a) | Vicuna-7B | 63.3 | 68.4 | 68.8 | 62.2 | 52.0 | 16.8 |
| LLaVA-1.5 (Liu et al., 2024) | Vicuna-7B | 62.4 | 69.0 | 64.3 | 58.3 | 44.3 | **17.0** |
| LLaVA-1.5 + **MoDA (ours)** | Vicuna-7B | **62.5** | **71.0** | **64.8** | **58.6** | **53.4** | 13.2 |
| LLaVA-More OpenAI CLIP (Cocchi et al., 2025) | LLaMA 3.1-8B | 63.6 | 76.3 | **72.3** | **68.2** | 57.1 | 15.5 |
| LLaVA-More OpenAI CLIP **+ MoDA (ours)** | LLaMA 3.1-8B | **64.4** | **77.8** | 72.0 | 66.1 | **58.0** | **15.6** |
| LLaVA-More SigLIP-S2 (Cocchi et al., 2025) | LLaMA 3.1-8B | 64.9 | 77.1 | 71.8 | **68.0** | 57.2 | 17.3 |
| LLaVA-More SigLIP-S2**+ MoDA (ours)** | LLaMA 3.1-8B | **65.4** | **81.9** | **72.4** | 63.6 | **58.2** | **18.1** |

Table 2: **Performance on vision-centric benchmarks requiring fine-grained visual understanding.** We evaluate on LLaVA-Wild, MMVet, MMStar, V*Bench, and CV-Bench. **Bold underlined** values indicate highest scores per benchmark. **Bold** values show best performance within each baseline comparison. Gray text indicates models trained on different data distributions. All metrics are percentages; higher is better.

| Method | LLM | LLaVA-Wild | MMVet | MMStar | V*Bench | CV-Bench |
|---|---|---|---|---|---|---|
| BLIP-2 (Li et al., 2023a) | FLAN-T5 | 38.1 | - | 37.6 | - | - |
| InstructBLIP (Dai et al., 2023) | Vicuna-7B | 60.9 | 26.2 | 1.0 | 34.0 | - |
| Qwen-VL-Chat (Bai et al., 2023) | Qwen-7B | - | - | 37.7 | - | - |
| LLaVA (Liu et al., 2023a) | Vicuna-7B | 62.8 | 23.8 | - | 35.5 | - |
| LLaVA-1.5 (Liu et al., 2024) | Vicuna-13B | 72.5 | - | - | - | 60.9 |
| ShareGPT-4V (Chen et al., 2024a) | Vicuna-7B | 72.6 | - | 33.0 | - | 61.8 |
| LLaVA-1.5 (Liu et al., 2024) | Vicuna-7B | 65.4 | 28.1 | 27.6 | 42.9 | **59.0** |
| LLaVA-1.5 **+ MoDA (ours)** | Vicuna-7B | **68.0** | **29.9** | **32.9** | **44.5** | 58.2 |
| LLaVA-More OpenAI CLIP (Cocchi et al., 2025) | LLaMA 3.1-8B | 71.2 | 25.2 | 35.7 | 42.8 | 59.9 |
| LLaVA-More OpenAI CLIP **+ MoDA (ours)** | LLaMA 3.1-8B | **73.9** | **26.6** | **36.7** | **44.0** | **61.0** |
| LLaVA-More SigLIP-S2 (Cocchi et al., 2025) | LLaMA 3.1-8B | **72.0** | 27.7 | 35.8 | 44.4 | 61.2 |
| LLaVA-More SigLIP-S2 **+ MoDA (ours)** | LLaMA 3.1-8B | 67.6 | **28.3** | **38.5** | **44.8** | **62.2** |

## 4.1 RESULTS

We evaluate MoDA across 13 benchmarks spanning visual question answering, vision-centric reasoning, and hallucination detection. The overall trend aligns with our motivation (Section 1 and Section 3): by applying cross-attentive channel modulation, MoDA directs information flow toward instruction-relevant features and enables high-capacity encoders to produce more precise and well-grounded outputs.

Table 3: **Performance on hallucination detection benchmarks.** **Bold underlined** values indicate highest scores per benchmark. **Bold** values show best performance within each baseline comparison. Models marked with * use Gemma 3 (Team, 2025) as grader. All metrics are percentages; higher is better.

| Method | LLM | POPE | MMVP* |
|---|---|---|---|
| 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 (ours)** | Vicuna-7B | **87.1** | **36.0** |
| LLaVA-More OpenAI CLIP (Cocchi et al., 2025) | LLaMA 3.1-8B | 85.1 | 27.3 |
| LLaVA-More OpenAI CLIP **+ MoDA (ours)** | LLaMA 3.1-8B | **86.3** | **28.7** |
| LLaVA-More SigLIP-S2 (Cocchi et al., 2025) | LLaMA 3.1-8B | 86.0 | 39.3 |
| LLaVA-More SigLIP-S2 **+ MoDA (ours)** | LLaMA 3.1-8B | **87.7** | **42.7** |

**VQA Performance.** As shown in Table 1, MoDA improves VQA by transforming the instruction into a soft, channel wise mask over visual embeddings. The gains scale with encoder quality. With SigLIP S2, ScienceQA increases by 4.8 points, from 77.1 to 81.9, and MoDA attains the highest scores on five of six VQA benchmarks: GQA at 65.4, ScienceQA at 81.9, MMBench En at 72.4, RealWorldQA at 58.2, and ChartQA at 18.1. An unexpected outcome appears on MMBench Cn. Vicuna 7B benefits slightly, moving from 58.3 to 58.6, while OpenAI CLIP and SigLIP S2 regress, moving from 68.2 to 66.1 and from 68.0 to 63.6. This behavior is consistent with a training mix dominated by English instructions and suggests that multilingual instruction tuning should recover the advantage without modifying the mechanism. Importantly, this limitation also supports our design. The decrease indicates that MoDA relies on instruction language conditioning rather than on parameter count, since a pure capacity increase would likely raise scores across languages uniformly. This is straightforward to address by adding multilingual instructions during tuning, so we view it as a data coverage issue rather than a weakness of our method. On ChartQA, our scores were lower because the LLaVA-1.5 tuning set lacked plot/chart data, limiting exposure to visual chart reasoning.

**Vision Centric Tasks.** On the benchmarks that require careful visual discrimination, shown in Table 2, architectural precision outperforms parameter count and follows our motivation. Patch tokenization mixes multiple semantics inside each token. MoDA applies cross attentive, instruction conditioned channel modulation that separates useful signals from unrelated content and routes them more effectively to the decoder. This converts the representational headroom in stronger encoders into measurable accuracy. OpenAI CLIP with MoDA reaches the best LLaVA Wild score at 73.9. The peak on MMVet is achieved by the compact Vicuna 7B with MoDA at 29.9. SigLIP S2 with MoDA attains the strongest results on MMStar at 38.5, on V*Bench at 44.8, and on CV Bench at 62.2. These datasets emphasize different skills such as recognition, reading, and spatial reasoning, yet the pattern is consistent. The largest gains appear when MoDA is paired with SigLIP S2, which provides richer features that MoDA can selectively emphasize. Importantly, MoDA also competes with models trained on larger and different data distributions. ShareGPT 4V, reported in gray, records 72.6 on LLaVA Wild, 33.0 on MMStar, and 61.8 on CV Bench. MoDA surpasses these results with 73.9 on LLaVA Wild, 38.5 on MMStar, and 62.2 on CV Bench. Comparisons to 13B baselines, including ShareGPT 4V, indicate that an 8B class model with MoDA can meet or exceed larger systems where direct comparisons exist. This favors design choices that direct information flow over simply adding parameters and matches the behavior predicted by the method.

**Hallucination Detection.** MoDA's design intent is most evident on hallucination benchmarks, as summarized in Table 3. By emphasizing instruction relevant channels and attenuating distractors, the model reduces reliance on priors and keeps outputs consistent with the visible content. With Vicuna 7B, MMVP improves by 12.0 points, from 24.0 to 36.0. With SigLIP S2, MoDA attains the top scores on both tasks, reaching 87.7 on POPE and 42.7 on MMVP, and surpasses the 13B LLaVA 1.5 baseline, which records 85.9 on POPE and 24.7 on MMVP. Taken together, the results confirm three discoveries. First, MoDA scales with stronger encoders, most clearly with SigLIP S2. Second, architectural refinement yields larger benefits than parameter growth in multiple settings. Third,

Table 4: **Ablation Study of MoDA Components.** We systematically evaluate MoDA architecture variants (Linear MLP vs. Cross-Attention vs. Self-Attention), auxiliary supervision ($L_1$ vs. None), LLM backbones (Vicuna-7B vs. LLaMA 3.1-8B), and vision encoders (CLIP vs. SigLIP-S2). Cross-Attention without auxiliary loss consistently outperforms alternatives, with benefits amplified by stronger visual encoders. Bold values indicate best performance per column.

| 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** |

hallucination detection is where MoDA delivers its most decisive gains. Across all three categories, MoDA achieves the best result on 12 of the 13 benchmarks. These gains are consistent with the mechanism described in Section 3, where instruction conditioned channel modulation reduces the influence of mixed patch semantics. The improvements require no additional supervision or changes to the training protocol, indicating that MoDA improves how existing evidence is used rather than expanding data or labels.

### 4.2 Ablation Studies

We conduct systematic ablations to address key reviewer concerns: **(i)** Why Cross-Attention outperforms linear modulation, **(ii)** Whether improvements stem from architecture vs. added capacity, **(iii)** Component synergy effects across different encoders and LLMs.

**Cross-Attention vs. Alternatives:** To understand why Cross-Attention outperforms alternatives, we analyze how each approach handles queries requiring disentangling mixed visual semantics within individual patches. The three approaches differ fundamentally: Linear MLP applies the same transformation regardless of instruction, Self-Attention concatenates features without explicit cross-modal conditioning, while Cross-Attention uses visual features as queries and instruction tokens as memory, enabling selective channel emphasis based on instruction semantics. This architectural difference becomes crucial when processing patches containing multiple semantic elements, as Cross-Attention can dynamically weight channels corresponding to instruction-relevant concepts while suppressing irrelevant information. With SigLIP-S2, Cross-Attention achieves the highest performance (69.4 vs Self-Attention 68.0 vs Linear 67.3) with substantial gains on reasoning tasks: ScienceQA shows Cross-Attention at 81.9 compared to Self-Attention 79.9 and Linear 77.8, while MMVP demonstrates Cross-Attention's 42.7 versus Self-Attention's 39.5 and Linear's 40.0.

**Architecture vs. Capacity:** The performance patterns argue against pure capacity effects: task-specific rather than uniform improvements (MMVP shows large gains while other tasks show smaller improvements), consistent improvement patterns across different LLM backbones, and architectural choice matters more with stronger components (differences are minimal with CLIP but substantial with SigLIP-S2).

**Component Synergy:** $L_1$ regularization consistently degrades Cross-Attention performance across both encoders, while Linear MLP remains largely unaffected. The degradation is particularly severe for fine-grained reasoning. LLaMA 3.1-8B provides modest improvements over Vicuna-7B, while SigLIP-S2 dramatically amplifies MoDA's effectiveness (+5.1 points over CLIP), confirming that instruction-guided modulation becomes increasingly valuable with richer visual representations.

**Additional Analysis.** Appendix ablations validate MoDA's depth (A.3.1), placement (A.3.2), and qualitative performance (A.4), highlighting its fine-grained understanding. See Appendix for details.

Table 5: **Comparison with masking approaches.** We compare MoDA against token-level masking methods using identical conditions (LLaMA 3.1-8B + SigLIP-S2). **Bold** values indicate best performance per column.

| Strategy | POPE | GQA | SQA | MMVP | Avg |
|---|---|---|---|---|---|
| Baseline | 86.0 | 64.9 | 77.1 | 39.3 | 66.8 |
| Learnable Masking (Barrios & Jin, 2024) | 86.9 | 65.1 | 79.9 | 41.9 | 68.5 |
| Sparse Masking (Lin et al., 2022) | 85.8 | 64.7 | 76.7 | 38.8 | 66.5 |
| **MoDA (ours)** | **87.7** | **65.4** | **81.9** | **42.7** | **69.4** |

### 4.3 Comparison with Masking Approaches

Table 5 validates our core hypothesis by comparing MoDA against token-level masking methods under identical conditions. MoDA achieves the highest performance across all benchmarks, establishing clear superiority with 69.4 average performance compared to 68.5 for Learnable Masking and 66.5 for Sparse Masking. Most importantly, MoDA reaches the strongest results on fine-grained tasks: 42.7 on MMVP and 81.9 on ScienceQA. While token-level masking operates on discrete attention weights and requires layer-wise computation scaling with model depth, MoDA's channel-wise modulation provides continuous, instruction-guided refinement with single-pass efficiency, enabling more effective visual-language understanding without computational overhead that increases linearly with the number of transformer layers.

Table 6: **Computational overhead of MoDA relative to LLaVA-MoRE.** MoDA introduces minimal overhead with only 3.7% of total parameters and less than 1% of computational operations (MACs and FLOPs), showing that performance gains stem from architectural innovation rather than scaling.

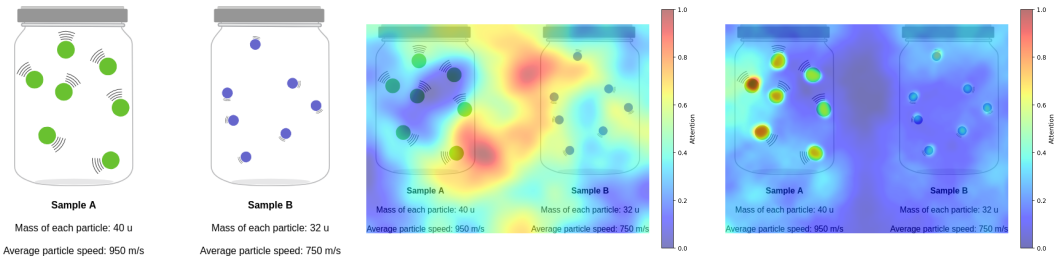
| Metric | MoDA | LLaVA-MoRE (8B) | Ratio (%) |
|---|---|---|---|
| Parameters | 0.302B | 8.0B | 3.7 |
| MACs | 45.1G | $\approx$ 5,246G | 0.86 |
| FLOPs | 90.2G | $\approx$ 10,492G | 0.86 |

### 4.4 Computational Efficiency Analysis

MoDA introduces minimal overhead, adding only 3.7% parameters and $<$1% MACs/FLOPs compared to LLaVA-MoRE (8B) (Table 6), confirming gains stem from architectural design rather than scaling. MoDA's strategic placement after the adapter and before the LLM enables instruction-guided modulation with optimal efficiency-performance tradeoffs, as validated by our ablation studies comparing different placement strategies (Appendix A.3.2). This positioning allows MoDA to operate on pre-aligned visual features while maintaining computational efficiency. In multi-turn scenarios, visual features are cached once, with subsequent queries requiring only modulation recomputation ($<$1% computation).

### 4.5 Attention Map Visualization

To provide insight into how MoDA improves visual grounding, we visualize attention maps derived from the LLM's self-attention layers. In MLLMs, visual tokens are concatenated with language tokens and processed jointly through the transformer layers. We extract the attention weights from the




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

Figure 3: **Attention map visualization on ScienceQA.** Given the question *"Which sample has the higher temperature?"*, the baseline model (b) exhibits diffuse attention across both containers and irrelevant regions, leading to an incorrect response. In contrast, MoDA (c) concentrates attention on Sample A's particles and motion indicators, enabling the model to correctly identify Sample A as having higher temperature.

output token positions attending to the visual token positions, then spatially reshape these weights to match the original image resolution. Figure 3 presents a representative example from ScienceQA, where the task requires comparing the average kinetic energies of gas particles across two containers to determine which sample exhibits higher temperature. The baseline model produces a diffuse attention distribution across both containers and irrelevant background regions, indicating that the LLM struggles to focus on task-relevant visual tokens, leading to an incorrect prediction. In contrast, when visual features are pre-processed through MoDA's channel-wise modulation, the attention maps exhibit concentrated activation patterns localized on Sample A's particles and their associated velocity indicators, which are directly relevant to solving the task. This demonstrates that MoDA's instruction-guided modulation effectively refines visual token representations, enabling the LLM to allocate attention more precisely to task-relevant regions. These visualizations provide interpretable evidence that channel-wise feature modulation enhances visual-language alignment, facilitating accurate fine-grained visual reasoning.

## 5 Conclusions

We have introduced MoDA a novel modulation adapter for MLLMs that works as an ad-hoc module. At its core, MoDA re-weights the contribution of each individual visual feature channel based on the early language embeddings of the language prompt. The re-weighted set of features acts as an implicit feature selector promoting the relevant visual features which are more relevant for each individual query, thus improving the performance of MLLMs. Across multiple benchmarks and multiple MLLM architectures MoDA shows consistent performance improvements over the baselines. MoDA does not require any additional pre-training or supervision. By simply appending MoDA to the MLLM during the instructional tuning phase, we observe direct improvements across diverse benchmarks.

**Limitations.** MoDA works by directly modulating the channels in the input, but it can not achieve explicit sparsity in the channel dimension. That is, MoDA re-weights the channel dimension but only occasionally it would set a channel's weight to 0. Such property could be desirable to make a stronger feature selection and effectively guide the attention of the LLM towards the more semantically relevant features.

**Reproducibility Statement.** We place the highest priority on reproducibility. Upon acceptance, we will release the MoDA MLLM model weights, along with the training pipeline, including all hyperparameter configurations. During experimentation we fixed the random seeds and explicitly set key parameters that control the variability of the underlying LLMs, for example, setting `do_sample=False`, to eliminate sources of stochasticity.

## REFERENCES


Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Wayner Barrios and SouYoung Jin. Multi-layer learnable attention mask for multimodal tasks, 2024. URL https://arxiv.org/abs/2406.02761.

Jinhe Bi et al. Llava steering: Visual instruction tuning with 500x fewer parameters through modality linear representation-steering. *arXiv preprint arXiv:2412.12359*, 2024.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XVII*, pp. 370–387, Berlin, Heidelberg, 2024a. Springer-Verlag. ISBN 978-3-031-72642-2. doi: 10.1007/978-3-031-72643-9_22. URL https://doi.org/10.1007/978-3-031-72643-9_22.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024b.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tai, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25(1), January 2024. ISSN 1532-4435.

Federico Cocchi, Nicholas Moratelli, Davide Caffagni, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. Llava-more: A comparative study of llms and visual backbones for enhanced visual instruction tuning, 2025. URL https://arxiv.org/abs/2503.15621.

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. URL https://api.semanticscholar.org/CorpusID:258615266.

Zhihao Fan, Yeyun Gong, Dayiheng Liu, Zhongyu Wei, Siyuan Wang, Jian Jiao, Nan Duan, Ruofei Zhang, and Xuanjing Huang. Mask attention networks: Rethinking and strengthen transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1692–1701, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.135. URL https://aclanthology.org/2021.naacl-main.135.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702, 2019. doi: 10.1109/CVPR.2019.00686.

Ouguzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. Brave: Broadening the visual encoding of vision-language models. In *European Conference on Computer Vision*, 2024. URL https://api.semanticscholar.org/CorpusID:269033274.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023a. URL https://api.semanticscholar.org/CorpusID:256390509.

others Li. Instruction-guided fusion of multi-layer visual features in large vision-language models. *arXiv preprint arXiv:2501.08443*, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji rong Wen. Evaluating object hallucination in large vision-language models. In *Conference on Empirical Methods in Natural Language Processing*, 2023b. URL https://api.semanticscholar.org/CorpusID:258740697.

Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, and Jinqiao Wang. Mst: Masked self-supervised transformer for visual representation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 13165–13176. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/6dbbe6abe5f14af882ff977fc3f35501-Paper.pdf.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2023. URL https://arxiv.org/abs/2211.09110.

Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *CVPR*, 2022.

Te Lin and Inwhee Joe. An adaptive masked attention mechanism to act on the local text in a global context for aspect-based sentiment analysis. *IEEE Access*, 11:43055–43066, 2023. doi: 10.1109/ACCESS.2023.3270927.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296, 2024. doi: 10.1109/CVPR52733.2024.02484.

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*, 2023b. URL https://api.semanticscholar.org/CorpusID:259837088.

Llama Team, AI @ Meta. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Jie Ma, Yalong Bai, Bineng Zhong, Wei Zhang, Ting Yao, and Tao Mei. Visualizing and understanding patch interactions in vision transformer. *IEEE Transactions on Neural Networks and Learning Systems*, 35:13671–13680, 2022. URL https://api.semanticscholar.org/CorpusID:247410956.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning, 2022. URL https://arxiv.org/abs/2203.10244.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

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, 2021. URL https://arxiv.org/abs/2103.00020.

Riccardo Rende, Federica Gerace, Alessandro Laio, and Sebastian Goldt. What does self-attention learn from masked language modelling?, 2024.

Rui Shi, Tianxing Li, Liguo Zhang, and Yasushi Yamaguchi. Visualization comparison of vision transformers and convolutional neural networks. *IEEE Transactions on Multimedia*, 26:2327–2339, 2024. doi: 10.1109/TMM.2023.3294805.

Jingfan Tang, Xinqiang Wu, Min Zhang, Xiujie Zhang, and Ming Jiang. Multiway dynamic mask attention networks for natural language inference. *J. Comp. Methods in Sci. and Eng.*, 21(1): 151–162, jan 2021. ISSN 1472-7978. doi: 10.3233/JCM-204451. URL https://doi.org/10.3233/JCM-204451.

Gemma Team. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024b. URL https://arxiv.org/abs/2401.06209.

Andrés Villa, Juan Carlos León Alcázar, Alvaro Soto, and Bernard Ghanem. Behind the magic, merlim: Multi-modal evaluation benchmark for large image-language models, 2024. URL https://arxiv.org/abs/2312.02219.

Andrés Villa, Juan León Alcázar, Motasem Alfarra, Vladimir Araujo, Alvaro Soto, and Bernard Ghanem. Eagle: Enhanced visual grounding minimizes hallucinations in instructional multimodal models, 2025. URL https://arxiv.org/abs/2501.02699.

Yaqing Wang et al. Non-intrusive adaptation: Input-centric parameter-efficient fine-tuning for versatile multimodal modeling. *arXiv preprint arXiv:2310.12100*, 2023.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark, 2024. URL https://arxiv.org/abs/2406.01574.

xAI. Grok-1.5 vision (grok-1.5v) preview: Connecting the digital and physical worlds with our first multimodal model. xAI News, April 2024. URL https://x.ai/news/grok-1.5v. Preview announcement.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2024. URL https://arxiv.org/abs/2308.02490.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. URL https://arxiv.org/abs/2303.15343.

Fan Zhang, Shulin Tian, Ziqi Huang, Yu Qiao, and Ziwei Liu. Evaluation agent: Efficient and promptable evaluation framework for visual generative models. *arXiv preprint arXiv:2412.09645*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=uccHPGDlao.

Tao Zhou, Yuxia Niu, Huiling Lu, Caiyue Peng, Yujie Guo, and Huiyu Zhou. Vision transformer: To discover the "four secrets" of image patches. *Information Fusion*, 105:102248, 2024. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2024.102248. URL https://www.sciencedirect.com/science/article/pii/S1566253524000265.

# A Appendix

## A.1 Experiment Setup

**Visual Question Answering Benchmarks.** These benchmarks evaluate models' capability to accurately answer questions requiring visual reasoning and comprehension. **GQA** (Hudson & Manning, 2019) builds upon Visual Genome's scene graph annotations and contains 113k images with 22 million questions emphasizing compositional reasoning and scene understanding. **ScienceQA** (Lu et al., 2022) assesses models using complex multimodal multiple-choice questions spanning three major subject areas (natural science, language science, and social science), encompassing 26 topics, 127 categories, and 379 distinct skills across 4,241 test examples. **MMBench** (Liu et al., 2023b) consists of approximately 3,000 multiple-choice questions spanning 20 diverse domains, designed to rigorously assess MLLM capabilities across perception and reasoning paradigms through a structured hierarchical taxonomy. We evaluate on both English (MMBench-En) and Chinese (MMBench-Cn) versions to assess multilingual capabilities. **RealWorldQA** (xAI, 2024) contains over 700 real-world images captured from vehicles and other scenarios, each paired with spatial reasoning questions that evaluate real-environment understanding and physical scene comprehension. **ChartQA** (Masry et al., 2022) focuses on chart understanding with 9.6k human-written and 23k auto-generated questions across approximately 20k charts (bar, line, pie), requiring visual and logical reasoning such as comparing values, identifying trends, and performing arithmetic operations over chart data.

**Vision-Centric Benchmarks.** These benchmarks specifically target fine-grained visual understanding and detailed image analysis capabilities. **LLaVA-Wild** (Liu et al., 2023a) comprises 24 diverse images with 60 questions spanning indoor and outdoor scenes, memes, paintings, and sketches, with each image paired with detailed, manually curated descriptions and targeted questions categorized into conversation, detailed description, and complex reasoning. **MM-Vet** (Yu et al., 2024) includes 200 test images with 218 questions covering six core vision-language capabilities: recognition, knowledge, optical character recognition (OCR), spatial awareness, language generation, and mathematics, often requiring integration of multiple skills for accurate responses. **MMStar** (Chen et al., 2024b) presents 1,500 manually curated multimodal challenge items with minimal overlap, evaluating six high-level capabilities across 18 fine-grained axes and targeting complex visual dependency and reasoning tasks where visual content is essential for answering. **V*Bench** (Zhang et al., 2024) evaluates detailed visual analysis using 191 high-resolution images from SAM with average resolution of 2246×1582, containing two sub-tasks: attribute recognition (115 samples requiring recognition of object attributes like color and material) and spatial relationship reasoning (76 samples asking for relative spatial relationships between objects). **CV-Bench** (Tong et al., 2024a) provides

a comprehensive evaluation framework with 2,638 manually-inspected examples, repurposing standard vision benchmarks such as ADE20K, COCO, and Omni3D to assess both 2D understanding (spatial relationships, object counting) and 3D understanding (depth order, relative distance) within a multimodal context.

**Hallucination Detection Benchmarks.** These benchmarks specifically measure model tendency to generate false or inconsistent information not present in the visual input. **POPE** (Li et al., 2023b) evaluates object hallucination through 8,910 binary classification queries across three subsets (random, popular, and adversarial), each constructed via distinct sampling strategies to probe different aspects of hallucination phenomena in MLLMs. Following standard practice, we report average performance across all three subsets. **MMVP** (Tong et al., 2024b) measures hallucination through 150 carefully constructed image pairs, each accompanied by two binary-choice questions. The image pairs are selected to have highly similar CLIP embeddings, and accurate performance requires both questions per pair to be answered correctly, making this benchmark particularly challenging for detecting subtle visual differences and avoiding spurious correlations.

### A.2 Implementation Details

Table S1 summarizes all optimization, hardware, and architectural specifications needed to reproduce our results. We followed LLaVA's (Liu et al., 2023a; 2024) established two-stage training curriculum. First, we train the adapters on $558$K alt-text image-caption pairs, then fine-tune the network on high-quality visual instruction data. Both stages optimize the same next-token prediction objective, allowing us to maintain optimizer state and the cosine learning rate schedule with $3\%$ warm-up across stages.

Table S1: **Training Configuration Summary.** This table details the full set of training and fine-tuning hyper-parameters used to replicate our experimental setup. "PT" refers to the pre-training stage using large-scale alt-text image–caption data, while "FT" denotes the fine-tuning stage on high-quality visual-instruction datasets. Parameters are organized across optimization settings, hardware, and model architecture components shared across both stages.

| **Hyper-parameter** | **PT** | **FT** |
|---|---|---|
| Global batch size | 256 | 128 |
| Effective epochs | 1 | 1 |
| Learning rate | $1 \times 10^{-3}$ | $2 \times 10^{-5}$ |
| LR schedule | Cosine decay with 3 % warm-up | |
| Weight decay | 0 | |
| Optimiser | AdamW | |
| DeepSpeed stage | 2 | 3 |
| *Hardware* | | |
| GPU type | A100/H100 (80 GB) | |
| Deployment | Multi-node cluster | |
| *Model components (shared across stages)* | | |
| Language backbone | LLaMA-3.1-8B or Vicuna-7B | |
| Visual encoder | CLIP or SigLIP with S2 multiscale | |
| Adapter (MoDA) | 2 × cross-attention; 16 heads | |

To ensure direct comparability, we matched all hyper-parameters (batch sizes, learning rates, weight decay, and optimizer choice) exactly as reported in LLaVA-1.5. Training was distributed across multi-node clusters using 80 GB A100 or H100 GPUs with DeepSpeed ZeRO Stage-2 for pre-training and Stage-3 for fine-tuning. The model architecture combines either LLaMA-3.1-8B (Llama Team, AI @ Meta, 2024) or Vicuna-7B (Zheng et al., 2023) as the language backbone with CLIP (Radford et al., 2021) or SigLIP (Zhai et al., 2023) image encoders. Visual and textual in-

Table S2: **Ablation on MoDA Depth.** Effect of increasing the number of layers in the MoDA adapter while keeping every other component fixed. Scores are reported on POPE, GQA, SQA and MMVP; the final column shows the mean across tasks.

| MoDA type | # layers | Supp. loss | LLM | Vision enc. | POPE | GQA | SQA | MMVP | Avg |
|---|---|---|---|---|---|---|---|---|---|
| Linear (MLP) | 2 | None | LLaMA 3.1-8B | CLIP ViT-L/14@336 | **86.6** | **64.4** | **77.8** | **28.1** | **64.2** |
| Linear (MLP) | 4 | None | LLaMA 3.1-8B | CLIP ViT-L/14@336 | 82.0 | 57.7 | 42.1 | 27.3 | 52.3 |

Table S3: **Impact of the spatial reach of MoDA**. Comparison of LLaVA-More 8B without MoDA, with MoDA injected at the beginning of the LLM module, and with MoDA applied to every block of the LLM module. Scores are reported on POPE and MMVP (hallucination robustness), ScienceQA (scientific reasoning), and GQA (real-world visual reasoning); the final column shows the mean across tasks.

| Model | LLM size | LLM | MoDA position | Vision enc. | POPE | GQA | SQA | MMVP | Avg |
|---|---|---|---|---|---|---|---|---|---|
| LLaVA-More 8B | 8B | LLaMA 3.1-8B | - | SigLIP-S2 | 86.0 | 64.9 | 77.1 | 39.3 | 66.8 |
| LLaVA-More 8B + MoDA | 8B | LLaMA 3.1-8B | All layers in LLM | SigLIP-S2 | 86.3 | 65.1 | 78.9 | 39.8 | 67.5 |
| LLaVA-More 8B + MoDA | 8B | LLaMA 3.1-8B | Beginning | SigLIP-S2 | **87.7** | **65.4** | **81.9** | **42.7** | **69.4** |

formation merge through a two-layer MoDA cross-attention block where visual tokens query instruction embeddings.

**MMVP Benchmark.** To evaluate performance on the MMVP benchmark, we opted for an open-source and cost-effective alternative to proprietary language models. Instead of using GPT-4, we employed Gemma 3 (Team, 2025), as the grader (use only pure text). This model was deployed using `Ollama`, which ensures compatibility with the OpenAI API. This setup allowed us to maintain seamless integration with our Python-based evaluation pipeline while significantly reducing operation costs without compromising evaluation consistency.

## A.3 Additional Ablation Studies

### A.3.1 Effect of MoDA Adapter Depth

Table S2 showcases the impact of the adapter depth across four different evaluation protocols: POPE and MMVP, which target hallucination robustness; ScienceQA (SQA), which probes scientific reasoning; and GQA, a dataset for real world visual reasoning and compositional question answering. Not that the first row mirrors line 5 of Table 2 in the main paper. When we increase the MLP depth from two to four layers the average score falls by nearly twelve points with the largest drops on GQA and ScienceQA, suggesting that extra layers hinder the model's ability to align visual evidence with language semantics. We also do not observe any improvement in hallucination tests using POPE and MMVP. This indicates that deeper adapters add complexity without strengthening actual grounding. In short, with the current data regime increasing depth does not improve understanding, and MoDA with two layers remains the clear choice for balancing multimodal alignment, reasoning precision and resistance to hallucination.

### A.3.2 Influence of MoDA Modulation Depth

Table S3 indicates that increasing the depth of visual modulation does not invariably lead to superior performance. Introducing MoDA exclusively at the beginning of the language model raises the average score from $66.8$ to $69.4$, an improvement of $+2.6$ points. By comparison, extending MoDA to all transformer layers yields only $+0.7$ points, with the mean rising to $67.5$.

Examining individual benchmarks, the shallow configuration (at the beginning of the LLM module) attains the largest gains: $+1.7$ on POPE, $+0.5$ on GQA, $+4.8$ on ScienceQA, and $+3.4$ on MMVP relative to the baseline. The full-depth variant does not match these improvements across any task. The computational cost further accentuates this disparity. Employing MoDA at every transformer

block increases training time from approximately 20 hours to more than 50 hours. In contrast, the single-block alternative maintains the original computational budget.

**Takeaway.** Deploying MoDA at the first transformer block yields the most favourable balance between effectiveness and efficiency. This shallow configuration raises the mean accuracy from 66.8 to 69.4 (+2.6), while preserving the original training budget of roughly 20 hours. In contrast, distributing MoDA across all layers lifts the mean by only +0.7 points, yet extends training time beyond 50 hours. Hence, full-depth MoDA is justified only when marginal accuracy gains warrant a more than two-fold increase in computational cost; otherwise, a single MoDA layer remains the recommended default.

### A.4 Qualitative Analysis

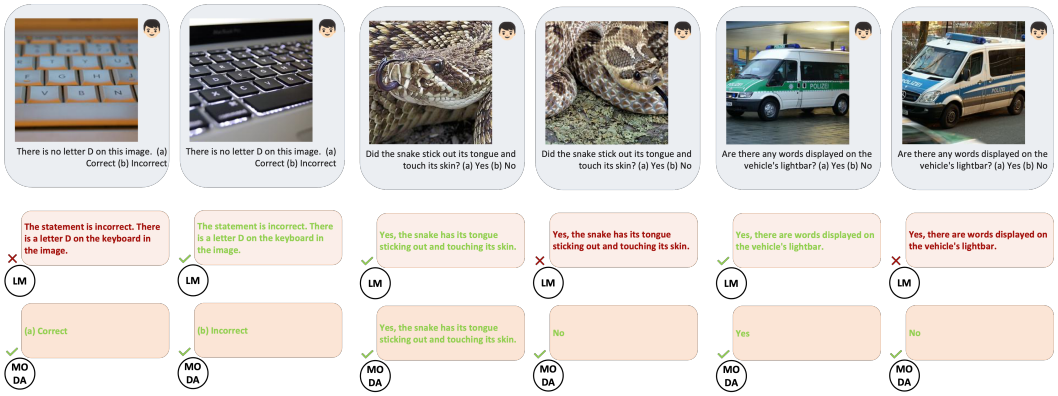



Figure S1: **Qualitative Analysis.** Qualitative comparison between the baseline LLaVA-More SigLIP-S2 (denoted **LM**) and LLaVA-More SigLIP-S2 + MoDA (denoted **MoDA**). Each column shows a multiple-choice VQA instance from the MMVP benchmark. ✗ marks an incorrect prediction, whereas ✔ denotes a correct one. Although the baseline frequently produces lengthy free-form answers that do not match the question format, MoDA consistently selects the correct alternative, successfully addressing the task. From left to right, we observe: (**i** & **ii**) recognition of a specific keyboard key, (**iii** & **iv**) detection of subtle tongue–skin contact in a snake, and (**iv** & **v**) identification of printed text on a police vehicle's light bar. Across all examples, MoDA demonstrates superior fine-grained grounding of visual cues.

Figure S1 presents a qualitative comparison between the baseline LLaVA-More using SigLIP-S2 (denoted as LM) and our proposed MoDA, which augments the same baseline with a modulation adapter to enhance visual representation quality. The first two examples involve determining whether the letter *D* is present in a keyboard layout. In the first case, LM incorrectly identifies the presence of the letter *D* despite its absence in the image and fails to select a valid multiple-choice option, resulting in both an incorrect response and invalid output format. In the second case, LM correctly identifies the letter's presence and selects the appropriate answer. In contrast, MoDA consistently selects the correct alternatives: "(a) Correct" for the first example and "(b) Incorrect" for the second, demonstrating its ability to produce concise outputs that comply with the required format while maintaining fine-grained visual understanding.

In the third example, both the baseline and MoDA produce correct answers but fail to follow the multiple-choice format. The fourth case involves identifying whether a snake's tongue is touching its skin, a subtle perceptual task where the correct answer is "No". LM misclassifies this contact while MoDA provides the correct answer, demonstrating greater sensitivity to localized visual features (fine-grained details).

The fifth and sixth examples test whether textual markings are present on a police van's light bar. In the fifth case, both models provide correct answers, but only MoDA follows the required output format instructions. In the sixth example, the LM incorrectly predicts the presence of text, likely due to overgeneralized visual priors, which are assumptions (hallucinations) formed from pre-training data that cause the model to expect text in similar visual contexts even when none is present. In contrast, MoDA accurately identifies the absence of text and maintains proper formatting. These

results highlight MoDA's improved grounding in visual evidence and its stronger compliance with formatting requirements, closely following the user's instructions.

### A.5 The Use of Large Language Models (LLMs).

We used commercial large language models (e.g., ChatGPT) only as editorial tools to improve the manuscript's readability. Their role was limited to language editing, such as correcting grammar, improving clarity, and smoothing the flow of text, and they did not influence the research design, data analysis, or the research conclusions.