# VISUAL INSTRUCTION TUNING WITH 500X FEWER PARAMETERS THROUGH MODALITY LINEAR REPRESENTATION-STEERING

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

## ABSTRACT

Multimodal Large Language Models (MLLMs) have significantly advanced visual tasks by integrating visual representations into large language models (LLMs). The textual modality, inherited from LLMs, equips MLLMs with abilities like instruction following and in-context learning. In contrast, the visual modality enhances performance in downstream tasks by leveraging rich semantic content, spatial information, and grounding capabilities. These intrinsic modalities work synergistically across various visual tasks. Our research initially reveals a persistent imbalance between these modalities, with text often dominating output generation during visual instruction tuning. This imbalance occurs when using both full fine-tuning and parameter-efficient fine-tuning (PEFT) methods. We then found that re-balancing these modalities can significantly reduce the number of trainable parameters required, inspiring a direction for further optimizing visual instruction tuning. Hence, in this paper, we introduce Modality Linear Representation-Steering (MoReS) to achieve the goal. MoReS effectively re-balances the intrinsic modalities throughout the model, where the key idea is to steer visual representations through linear transformations in the visual subspace across each model layer. To validate our solution, we composed LLaVA Steering, a suite of models integrated with the proposed MoReS method. Evaluation results show that the composed LLaVA Steering models require, on average, 500 times fewer trainable parameters than LoRA needs while still achieving comparable performance across three visual benchmarks and eight visual question-answering tasks. Last, we present the LLaVA Steering Factory, an in-house developed platform that enables researchers to quickly customize various MLLMs with component-based architecture for seamlessly integrating state-of-the-art models, and evaluate their intrinsic modality imbalance. This open-source project enriches the research community to gain a deeper understanding of MLLMs.

## 1 INTRODUCTION

Recent advancements in Multimodal Large Language Models (MLLMs) (Liu et al., 2024b; Xue et al., 2024; Zhou et al., 2024a; Chen et al., 2023) have demonstrated impressive capabilities across a variety of visual downstream tasks. These models integrate visual representations from pretrained vision encoders via various connectors (Liu et al., 2024a; Li et al., 2023a; Alayrac et al., 2022) into LLMs, leveraging the latter's sophisticated reasoning abilities (Zhang et al., 2024; Abdin et al., 2024; Zheng et al., 2023a).

To better integrate visual representations into LLMs, the most popular MLLMs adopt a two-stage training paradigm: pretraining followed by visual instruction tuning. In the pretraining stage, a connector is employed to project visual representations into the textual representation space. We define these two modalities—text and vision—as intrinsic to MLLMs, each carrying rich semantic information that serves as the foundation for further visual instruction tuning on downstream tasks

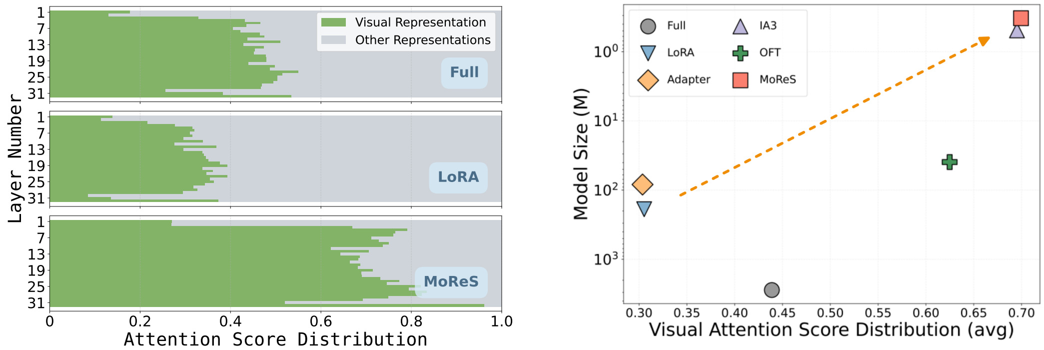


Figure 1: **Left:** Attention score distributions across layers for three MLLM fine-tuning methods (Full, LoRA, and MoReS), sampled from 100 instances each. Green represents visual representations, while grey indicates other (primarily textual) representations. Full fine-tuning and LoRA show strong reliance on textual representations across most layers. In contrast, the proposed MoReS method demonstrates significantly improved visual representation utilization, particularly in the middle and lower layers, addressing the intrinsic modality imbalance in MLLMs. **Right:** Average visual attention score distribution versus model size for different MLLM fine-tuning methods. The plot suggests that methods achieving better balanced intrinsic modality tend to require fewer trainable parameters.

such as image understanding (Sidorov et al., 2020), visual question answering (Goyal et al., 2017a; Lu et al., 2022; Hudson & Manning, 2019), and instruction following (Liu et al., 2023).

In the visual instruction tuning stage, due to its high computational cost, researchers have pursued two primary strategies. One approach focuses on refining data selection methodologies (Liu et al., 2024c; McKinzie et al., 2024) to reduce redundancy and optimize the training dataset, though this process remains expensive and time-consuming. A more common strategy goes to employ Parameter-Efficient Fine-Tuning (PEFT) methods, such as LoRA (Hu et al., 2021), aiming to reduce the number of trainable parameters, thereby making visual instruction tuning more computationally feasible (Liu et al., 2024a; Zhou et al., 2024a). However, even with PEFT methods like LoRA, large-scale MLLMs remain prohibitively expensive to fine-tuning.

This raises a critical question: is there any further possibility to reduce more trainable parameters so that the visual instruction tuning can be further improved? Our research offers a novel viewpoint by focusing on the intrinsic modality imbalance within MLLMs. A closer analysis uncovers an imbalance in output attention computation (Chen et al., 2024a), where textual information tends to dominate the attention distribution during output generation. Specifically, we investigate this issue by analyzing attention score distributions, which evaluates the balance between text and visual modalities. As shown in Figure 1, visual representations are significantly underutilized during visual instruction tuning. More importantly, our analysis reveals that achieving a better balance between these modalities can substantially reduce the number of trainable parameters required for fine-tuning. Hereby we suppose that *intrinsic modality rebalance is the Midas touch to unlock further reductions in the number of trainable parameters.*

To address this challenge, we introduce Modality Linear Representation-Steering (MoReS) to optimize visual instruction tuning, significantly reducing the number of trainable parameters while maintaining equivalent performance. Unlike full fine-tuning, which modifies the entire model, or other popular PEFT methods such as LoRA (Hu et al., 2021), OFT (Qiu et al., 2023), Adapter (Houlsby et al., 2019), and IA3 (Liu et al., 2022), MoReS focuses solely on steering the visual representations. Specifically, our approach freezes the entire LLM during visual instruction tuning to preserve its capabilities in the textual modality. Instead of fine-tuning the full model, we introduce a simple linear transformation to steer visual representations in each layer. This transformation operates within a subspace after downsampling, where visual representations encode rich semantic information in a compressed linear subspace (Zhu et al., 2024; Shimomoto et al., 2022; Yao et al., 2015). By continuously steering visual representations across layers, MoReS effectively controls the output generation process, yielding greater attention inclined to visual modality.

To validate the efficacy of our proposed MoReS method, we integrated it into MLLMs of varying scales (3B, 7B, and 13B parameters) during visual instruction tuning, following the LLaVA 1.5 (Liu et al., 2024a) training recipe. The resulting models, collectively termed LLaVA Steering, achieved competitive performance across three visual benchmarks and six visual question-answering tasks, while requiring 287 to 1,150 times fewer trainable parameters than LoRA, depending on the specific training setup.

In our experiments, we observed the need for a comprehensive framework to systematically analyze and compare various model architectures and training strategies in MLLMs. The wide range of design choices and techniques makes it difficult to standardize and understand the interplay between these components. Evaluating each method across different open-source models is time-consuming and lacks consistency due to implementation differences, requiring extensive data preprocessing and careful alignment between architectures and training recipes. To address this issue, we developed the LLaVA Steering Factory, a flexible framework that reimplements mainstream vision encoders, multi-scale LLMs, and diverse connectors, while offering customizable training configurations across a variety of downstream tasks. This framework simplifies pretraining and visual instruction tuning, minimizing the coding effort. Additionally, we have integrated our attention score distribution analysis into the LLaVA Steering Factory, providing a valuable tool to the research community for further studying intrinsic modality imbalance in MLLMs.

Our work makes the following key contributions to the field of MLLMs:

1. First of all, we propose Modality Linear Representation-Steering (MoReS), a novel method that addresses intrinsic modality imbalance in MLLMs by steering visual representations through linear transformations within the visual subspace, effectively mitigating the issue of text modality dominating visual modality.
2. In addition, we present LLaVA Steering, where with different sizes (3B/7B/13B), three real-world LLaVA MLLMs consisting of different model components are composed by integrating the proposed MoReS method into visual instruction tuning. LLaVA Steering models based on MoReS method achieve comparable performance across three visual benchmarks and six visual question-answering tasks, while requiring $287$ to $1,150$ times fewer trainable parameters.
3. Last but not least, we develop the LLaVA Steering Factory, a flexible framework designed to streamline the development and evaluation of MLLMs with minimal coding effort. It offers customizable training configurations across diverse tasks and incorporates tools such as attention score analysis, facilitating systematic comparisons and providing deeper insights into intrinsic modality imbalance.

## 2 Related Work

**Integrating Visual Representation into LLMs:** To leverage pre-trained large language models (LLMs) for understanding visual instructions and generating responses, researchers have introduced cross-attention mechanisms to integrate image information into the language model. Notable examples include models such as LLaMA 3-V (Dubey et al., 2024), IDEFICS (Laurençon et al., 2023), and Flamingo (Awadalla et al., 2023; Alayrac et al., 2022). These models typically follow a two-stage training process: pretraining on large-scale image-text datasets, followed by supervised fine-tuning (SFT) with carefully curated high-quality data. During this process, the self-attention layers in the LLM decoder are kept frozen, with only the cross-attention and perceiver layers updated, ensuring that the text-only performance remains intact.

Another prominent approach employs a decoder-only architecture, as seen in models like the LLaVA family (Liu et al., 2024b;a; 2023), BLIP (Xue et al., 2024; Li et al., 2023a), and Qwen-VL (team, 2024; Bai et al., 2023). These models also follow the pretraining and visual instruction tuning paradigm. In the pretraining stage, a randomly initialized connector is trained while keeping the LLM frozen. However, recent studies (Bai et al., 2023; Chen et al., 2023) have demonstrated scenarios where both the projector and vision encoder are jointly trained during pretraining. Given the limited capacity of adapter modules, it is common to unfreeze the LLM during visual instruction tuning, while keeping the vision encoder frozen.

NVLM (Dai et al., 2024) represents a hybrid approach, combining elements of both the cross-attention and decoder-only architectures. In contrast, vision-encoder-free methods, as explored by models like Fuyu (Bavishi et al., 2023), SOLO (Chen et al., 2024b), and EVE (Diao et al., 2024), directly integrate visual information into LLMs at the pixel level, foregoing traditional vision encoders altogether.

While these approaches have advanced the integration of visual representations into LLMs, they still face significant challenges in the computational demands of visual instruction tuning, motivating further exploration into more efficient methods.

**Visual Instruction Tuning:** Fine tuning of multimodal large language models (MLLMs) for downstream tasks has gained considerable attention, but remains computationally expensive due to large-scale visual instruction datasets and model sizes (Wang et al., 2022). To tackle this challenge, recent advancements have introduced parameter-efficient fine-tuning (PEFT) methods (Houlsby et al., 2019; Li & Liang, 2021), such as LoRA (Hu et al., 2021), enabling more efficient visual instruction tuning.

However, many of these PEFT methods primarily focus on optimizing weights but ignore the intrinsic representation imbalance during visual instruction tuning, thus cannot further reduce the required trainable parameters. This means to look for other novel approaches that can improve the efficiency and effectiveness of visual instruction tuning.

**Representation Steering:** Recent studies (Singh et al., 2024; Avitan et al., 2024; Li et al., 2024; Subramani et al., 2022) have demonstrated that the representations induced by pre-trained language models (LMs) encode rich semantic structures. Steering operations within this representation space have shown to be effective in controlling model behavior. Unlike neuron-based or circuit-based approaches, representation steering manipulates the representations themselves, providing a clearer mechanism for understanding and controlling the behavior of MLLMs and LLMs. For example, (Zou et al., 2023) explores representation engineering to modify neural network behavior, shifting the focus from neuron-level adjustments to transformations within the representation space. Similarly, (Wu et al., 2024a) applies scaling and biasing operations to alter intermediate representations. Furthermore, (Wu et al., 2024b) introduces a family of representation-tuning methods that allows for interpretable interventions within linear subspaces.

In this work, we leverage the concept of representation steering to introduce a novel approach, MoReS, which enhances attention to visual representations, thereby demonstrating superior parameter efficiency compared to baseline PEFT methods (Hu et al., 2021; Houlsby et al., 2019; Liu et al., 2022; Qiu et al., 2023).

## 3 INTRINSIC MODALITY IMBALANCE

This section explores how the two intrinsic modalities—text and vision—are imbalanced during output generation across each layer in MLLMs, as reflected in the attention score distribution. Furthermore, we demonstrate that addressing this modality imbalance effectively during visual instruction tuning can guide the design of methods that require fewer trainable parameters.

We begin with calculating the attention score distribution across both modalities in each layer, as derived from the generated output. In auto-regressive decoding, which underpins decoder-only MLLMs, output tokens are generated sequentially, conditioned on preceding tokens. The probability distribution over the output sequence $\hat{y}$ is formalized as:

$$p(\hat{y}) = \prod_{i=1}^{L} p(\hat{y}_i | \hat{y}_{<i}, R_{\text{text}}, R_{\text{image}}, R_{\text{sys}}) \tag{1}$$

where $\hat{y}_i$ represents the $i$-th output token, $\hat{y}_{<i}$ denotes the preceding tokens, $R_{\text{text}}$ is the textual representation, $R_{\text{image}}$ is the visual input representation, $R_{\text{sys}}$ accounts for system-level contextual information, and $L$ is the output sequence length.

To quantify modality representation imbalance, we calculate the sum of attention scores allocated to visual representations across all layers in MLLMs. Figure 1 illustrates this imbalance across full fine-tuning, LoRA, and our proposed MoReS method. The results indicate that textual representations often dominate the output generation process in both full fine-tuning and LoRA.

Further examination of this imbalance across multiple PEFT methods reveals an intriguing trend: methods that make better use of visual representations tend to require fewer trainable parameters during visual instruction tuning.

To validate this observation, we introduce the Layer-wise Modality Attention Ratio (LMAR), formulated as:

$$\mathrm{LMAR}_l = \frac{1}{N}\sum_{i=1}^{N}\frac{\alpha_l^{\text{image},i}}{\alpha_l^{\text{text},i}}, \tag{2}$$

where $l$ denotes the layer index, $N$ is the total number of samples, and $\alpha_l^{\text{image},i}$ and $\alpha_l^{\text{text},i}$ are the mean attention scores allocated to visual and textual tokens, respectively, in layer $l$ for the $i$-th sample. LMAR thus provides a robust measure of the attention distribution between modalities, averaged over multiple samples to capture general trends in modality representation across layers.

In our experiments comparing various existing PEFT methods and full fine-tuning, IA3 (Liu et al., 2022) consistently achieves the highest average LMAR score across all layers while requiring the fewest trainable parameters. IA3's superior performance can be attributed to its unique design, which introduces task-specific rescaling vectors that directly modulate key components of the Transformer architecture, such as the keys, values, and feed-forward layers.
Unlike methods that introduce complex adapters or fine-tune all parameters, IA3 optimizes a small but crucial set of parameters responsible for attention and representation learning. By applying element-wise scaling to the attention mechanisms, IA3 effectively re-balances the attention distribution across two intrinsic modalities. This design is particularly beneficial during visual instruction tuning, as it allows the model to dynamically reallocate more attention to visual representations without requiring many trainable parameters.

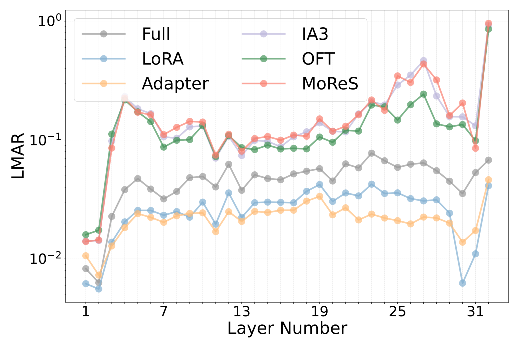


Figure 2: Layer-wise Modality Attention Ratio (LMAR) comparison across training methods, including Full fine-tuning, LoRA, Adapter, IA3, and our MoReS. Our MoReS method (red line) consistently demonstrates the highest LMAR across most layers, with a notable spike in the final layers. Compared with full fine-tuning and mainstream PEFT methods, our MoReS needs the least parameters during visual instruction tuning while achieving superior modality balance.

The identified relationship inspires that if the intrinsic modality imbalance can be addressed, the required number of trainable parameters can be potentially reduced further during visual instruction tuning. This offers a new direction for future improvements in PEFT methods for MLLMs.

## 4 MoReS Method

Based on insights gained from intrinsic modality imbalance, we introduce Modality Linear Representation-Steering (**MoReS**) as a novel method for visual instruction tuning which can rebalance visual and textual representations and achieve comparable performance with fewer trainable parameters.

Our approach is grounded in the linear subspace hypothesis, originally proposed by Bolukbasi et al. (2016), which suggests that information pertaining to a specific concept is encoded within a linear subspace in a model's representation space. This hypothesis has been rigorously validated across numerous domains, including language understanding and interpretability (Lasri et al., 2022; Nanda et al., 2023; Amini et al., 2023; Wu et al., 2024c).

Building upon the intervention mechanisms described in Geiger et al. (2024) and Guerner et al. (2023), we introduce a simple linear transformation that steers visual representations within subspace while keeping the entire LLM frozen during visual instruction tuning. This approach ensures that the language model's existing capabilities are preserved, while continuously guiding the MLLM to better leverage the underutilized visual modality. By steering visual representations across each layer, MoReS effectively rebalances the intrinsic modality and influences the output generation process. Figure 3 provides an illustration of the overall concept and architecture behind MoReS.

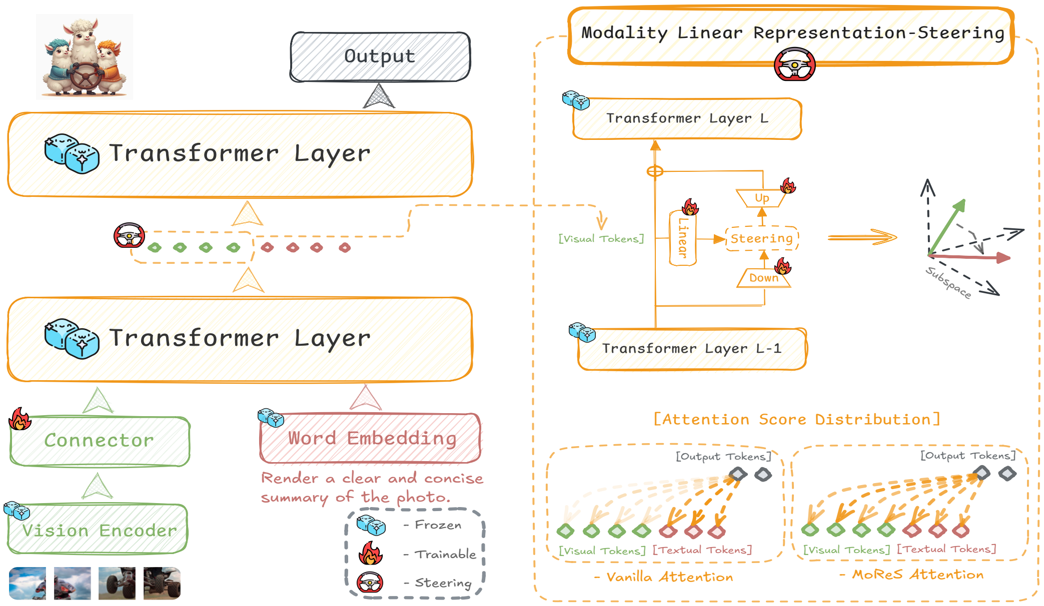


Figure 3: Schematic Overview of Modality Linear Representation-Steering (MoReS): **Left:** The architectural diagram depicts the integration of textual and visual tokens through transformer layers, leading to output token generation. **Right:** The mathematical formulation of MoReS illustrates the steering of visual representations within a subspace, highlighting its impact on output generation. During visual instruction tuning, the parameters of the LLM remain frozen, allowing only the parameters associated with the linear transformation in the steering mechanism to be trainable. With MoReS, the distribution of attention scores becomes more balanced, achieving intrinsic modality balance.

Formally, MoReS method can be formulated as follows: Let $\mathcal{H} = \{h_i\}_{i=1}^{N} \subset \mathbb{R}^D$ denote the set of visual representations in the original high-dimensional space. We define our steering function MoReS as:

$$\text{MoReS}(h) = W_{\text{up}} \cdot \phi(h) \tag{3}$$

where $h \in \mathbb{R}^D$ is an input visual representation, $\phi : \mathbb{R}^D \to \mathbb{R}^d$ is a linear transformation function that steers $h$ into a lower-dimensional subspace $\mathbb{R}^d$ ($d < D$), and $W_{\text{up}} \in \mathbb{R}^{D \times d}$ is an upsampling matrix that projects from $\mathbb{R}^d$ back to $\mathbb{R}^D$. The steering function $\phi$ is defined as:

$$\phi(h) = \text{Linear}(h) - W_{\text{down}} h \tag{4}$$

where $W_{\text{down}} \in \mathbb{R}^{d \times D}$ is a downsampling matrix. To preserve the fidelity of the representation and ensure a bijective mapping between spaces, we impose the following constraint $W_{\text{down}} W_{\text{up}}^T = I_D$. Notably, this steering method can dynamically be applied to specific visual tokens. Further exploration of the impact of different steered token ratios is discussed in Section 5.5.

In Section A.1, we further provide theoretical justification that elucidates how MoReS effectively rebalances the intrinsic modalities while continuously controlling output generation. Additionally, we provide a preliminary estimation of the trainable parameters involved during visual instruction tuning.

In the following sections, we first compose real-world MLLMs (i.e., LLaVA Steering) with three different scales and integrate the proposed MoReS method. Based on the composed real-world models, we then evaluate how our MoReS method performs within the composed models across several popular and prestigious datasets.

# 5 EXPERIMENTS

We incorporate MoReS into each layer of the LLM during visual instruction tuning, developing LLaVA Steering (3B/7B/13B) based on the training recipe outlined in (Liu et al., 2024a). During visual instruction tuning on the LLaVA-665k dataset, we apply MoReS to a specific ratio of the total visual tokens, specifically using it on only 1% of the tokens.

## 5.1 EXPERIMENT SETTINGS

### 5.1.1 LLAVA STEERING ARCHITECTURES

As illustrated in Figure 3, the architecture of the LLaVA Steering models (3B/7B/13B) consists of three essential components: a vision encoder, a vision connector responsible for projecting visual representations into a shared latent space, and a multi-scale LLM. The three modules are introduced below.

In our experiments, we utilize the Phi-2 2.7B model (Li et al., 2023c) alongside Vicuna v1.5 (7B and 13B) (Zheng et al., 2023b), sourced from our factory, to evaluate the generalizability of our approach across models of varying scales. For vision encoding, we employ CLIP ViT-L/14 336px (Radford et al., 2021) and SigLIP-SO400M-Patch14-384 (Zhai et al., 2023), while a two-layer MLP serves as the connector. Given the inefficiencies of Qformer in training and its tendency to introduce cumulative deficiencies in visual semantics (Yao et al., 2024), it has been largely replaced by more advanced architectures, such as the BLIP series (Xue et al., 2024), Qwen-VL series (team, 2024), and InternVL series (Chen et al., 2024c), which were previously reliant on Qformer.

### 5.1.2 BASELINE TRAINING METHODS

For comparison, four widely adopted PEFT methods (Adapter, LoRA, OFT and IA3) are selected as baselines. These methods establish a comparative framework to assess both the performance and efficiency of our proposed approach. Essentially, our MoReS method replaces these four PEFT methods during visual instruction tuning in LLaVA Steering.

**Adapter:** Building on the framework of efficient fine-tuning (Houlsby et al., 2019), we introduce adapter layers within Transformer blocks. These layers consist of a down-projection matrix $\mathbf{W}_{\text{down}} \in \mathbb{R}^{r\times d}$, a non-linear activation function $\sigma(\cdot)$, and an up-projection matrix $\mathbf{W}_{\text{up}} \in \mathbb{R}^{d\times r}$, where $d$ is the hidden layer dimension and $r$ is the bottleneck dimension. The adapter output is computed as:

$$\text{Adapter}(\mathbf{x}) = \mathbf{W}_{\text{up}}\sigma(\mathbf{W}_{\text{down}}\mathbf{x}) + \mathbf{x}, \tag{5}$$

where the residual connection ($+\mathbf{x}$) preserves the pre-trained model's knowledge. This formulation enables efficient parameter updates during fine-tuning, offering a balance between computational efficiency and adaptation capacity while minimally increasing the model's complexity.

**LoRA:** We employ the low-rank adaptation method (LoRA) proposed by (Hu et al., 2021), which efficiently updates the network's weights with a minimal parameter footprint by leveraging a low-rank decomposition strategy. For a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, the weight update is achieved through the addition of a low-rank decomposition, as shown in Equation 6:

$$W_0 + \Delta W = W_0 + BA \tag{6}$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are trainable low-rank matrices, and $r \ll \min(d, k)$.

**OFT:** We utilize the Orthogonal Finetuning (OFT) method, which efficiently fine-tunes pre-trained models by optimizing a constrained orthogonal transformation matrix (Qiu et al., 2023). For a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times n}$, OFT modifies the forward pass by introducing an orthogonal matrix $R \in \mathbb{R}^{d\times d}$, as illustrated in Equation 7:

$$z = W^\top x = (R \cdot W_0)^\top x \tag{7}$$

where $R$ is initialized as an identity matrix $I$ to ensure that fine-tuning starts from the pre-trained weights.

**IA3:** Building on the framework established by (Liu et al., 2022), we introduce three vectors $v_k \in \mathbb{R}^{d_k}$, $v_v \in \mathbb{R}^{d_v}$, and $v_{ff} \in \mathbb{R}^{d_{ff}}$ into the attention mechanism. The attention output is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q(v_k \odot K^T)}{\sqrt{d_k}}\right)(v_v \odot V), \tag{8}$$

where $\odot$ denotes multiplication by element.

### 5.2 Multi-Task Supervised Fine-tuning

To assess the generality of our method, we compare it with the baselines using the LLaVA-665K multitask mixed visual instruction dataset (Liu et al., 2024a). Our evaluation covers multiple benchmarks, including VQAv2 (Goyal et al., 2017b) and GQA (Hudson & Manning, 2019), which test visual perception through open-ended short answers, and VizWiz (Gurari et al., 2018), with 8,000 images designed for zero-shot generalization in visual questions posed by visually impaired individuals. We also use the image subset of ScienceQA (Lu et al., 2022) with multiple-choice questions to assess zero-shot scientific question answering, while TextVQA (Singh et al., 2019) measures performance on text-rich visual questions. MM-Vet (Yu et al., 2023) evaluates the model's ability to engage in visual conversations, with correctness and helpfulness scored by GPT-4. Additionally, POPE (Li et al., 2023b) quantifies hallucination of MLLMs. Finally, we apply the MMMU benchmark (Yue et al., 2024) to assess core multimodal skills, including perception, knowledge, and reasoning.

Following (Zhou et al., 2024b), we define ScienceQA as an unseen task, while VQAv2, GQA, and VizWiz are categorized as seen tasks in LLaVA-665k. To provide a comprehensive evaluation of our MoReS capabilities, we design three configurations: MoReS-Base, MoReS-Large, and MoReS-Huge, each based on different ranks.

We present the results in Table 1, where our MoReS method achieves the highest scores on POPE (88.2) and MMMU (35.8), as well as the second-best performance on ScienceQA (71.9) and MM-Vet (33.3). Notably, MoReS accomplishes these results with 287 to 1150 times fewer trainable parameters compared to LoRA. The scatter plots in Figure 4 further illustrate that MoReS variants (highlighted in red) consistently achieve Pareto-optimal performance, offering an ideal balance between model size and effectiveness.

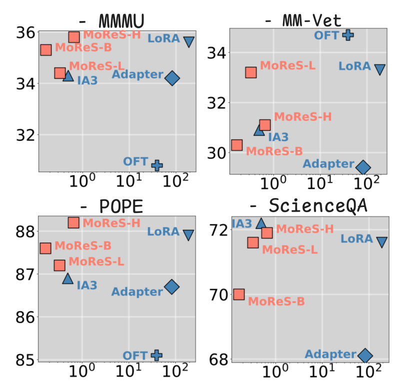


Figure 4: Comparison of parameter count vs. performance for MoReS and other PEFT methods across four benchmarks.

| Model | Method | TP* | VQAv2 | GQA | TextVQA | SciQA-IMG | POPE | MM-Vet | MMMU | Avg |
|---|---|---|---|---|---|---|---|---|---|---|
| LLaVA Steering-3B | FT | 2.78B | 79.2 | 61.6 | 57.4 | 71.9 | 87.2 | 35.0 | 38.2 | 61.5 |
| | Adapter | 83M | 77.1 | 58.9 | 53.5 | 68.1 | 86.7 | 29.4 | 34.2 | 58.2 |
| | LoRA | 188.74M | 77.6 | 59.7 | 53.8 | 71.6 | 87.9 | 33.3 | 35.6 | 59.9 |
| | OFT | 39.3M | 75.1 | 55.3 | 52.9 | 69.1 | 87.6 | 31.0 | 35.6 | 58.3 |
| | IA3 | 0.492M | 74.5 | 52.1 | 49.3 | 72.2 | 86.9 | 30.9 | 34.3 | 57.1 |
| | MoReS-B | 0.164M | 74.1 | 52.1 | 48.5 | 70.0 | 87.6 | 30.3 | 35.3 | 56.9 |
| | MoReS-L | 0.328M | 74.0 | 51.6 | 49.3 | 71.6 | 87.2 | 33.3 | 34.4 | 57.3 |
| | MoReS-H | 0.655M | 74.2 | 51.8 | 48.3 | 71.9 | 88.2 | 31.1 | 35.8 | 57.4 |

Table 1: Experimental results of Multi-Task Supervised Fine-tuning. For the TP* metric in this evaluation, we focus solely on the trainable parameters within the LLM. While different training strategies are applied to the vision encoder and connector across various recipes, we maintain a consistent training recipe for all models and benchmarks to ensure comparability

### 5.3 TASK-SPECIFIC FINE-TUNING

We evaluate the task-specific fine-tuning capabilities of our MoReS method in comparison to other tuning methods on multiple visual question answering datasets: (1) ScienceQA-Image (Lu et al., 2022), (2) VizWiz (Gurari et al., 2018), and (3) IconQA-txt and IconQA-blank (Lu et al., 2021).

We present the results in Table 2, showing that MoReS achieves 1200 times fewer trainable parameters compared to LoRA and 3 times fewer than the previous best, IA3, while maintaining comparable performance or an acceptable decline of less than 3%. These results show that MoReS can succeed at Task-Specific Fine-tuning, even unseen tasks during its multitask visual instruciton tuning stage.

| Model | Method | TP* | SciQA-IMG | VizWiz | IconQA-txt | IconQA-blank |
|---|---|---|---|---|---|---|
| LLaVA Steering-3B | Adapter | 83M | 92.3 | 62.9 | 93.5 | 95.8 |
| | LoRA | 188.7M | 93.9 | 61.6 | 93.9 | 96.5 |
| | OFT | 39.32M | 86.3 | 42.0 | 87.8 | 42.0 |
| | IA3 | 0.492M | 90.2 | 58.4 | 84.5 | 94.7 |
| | MoReS-B | 0.164M | 89.7 | 59.2 | 84.0 | 94.2 |
| LLaVA Steering-7B | Adapter | 201.3M | 82.7 | 59.7 | 72.1 | 71.6 |
| | LoRA | 319.8M | 87.6 | 60.6 | 77.7 | 70.2 |
| | OFT | 100.7M | 78.3 | 55.1 | 19.4 | 22.7 |
| | IA3 | 0.614M | 83.8 | 54.3 | 65.1 | 70.4 |
| | MoReS-B | 0.262M | 83.6 | 54.2 | 64.2 | 70.2 |
| LLaVA Steering-13B | Adapter | 314.6M | 87.9 | 61.4 | 78.2 | 73.0 |
| | LoRA | 500.7M | 92.1 | 62.0 | 80.2 | 73.2 |
| | OFT | 196.6M | 82.7 | 59.5 | 3.4 | 22.3 |
| | IA3 | 0.963M | 90.5 | 54.6 | 73.8 | 71.7 |
| | MoReS-B | 0.410M | 89.5 | 54.3 | 74.9 | 71.5 |

Table 2: Results of Task-Specific Fine-tuning, where higher values correspond to better performance.

| Scale | Method | TP* | SciQA-IMG | VizWiz | IconQA |
|---|---|---|---|---|---|
| Small | FT | 2.78B | 33.8 | 51.2 | 68.1 |
| | Adapter | 83M | 81.0 | 57.4 | 72.4 |
| | LoRA | 188.74M | 84.0 | 58.5 | 74.2 |
| | OFT | 39.32M | 79.2 | 43.2 | 35.9 |
| | IA3 | 0.492M | 79.9 | 50.5 | 73.0 |
| | MoReS-L | 0.328M | 78.2 | 55.0 | 69.7 |
| Medium | FT | 2.78B | 78.2 | 58.9 | 92.2 |
| | Adapter | 83M | 92.1 | 60.6 | 93.2 |
| | LoRA | 188.74M | 92.9 | 60.5 | 92.7 |
| | OFT | 39.32M | 86.4 | 44.4 | 45.5 |
| | IA3 | 0.492M | 91.9 | 57.1 | 90.6 |
| | MoReS-L | 0.328M | 92.1 | 56.6 | 89.9 |
| Large | FT | 2.78B | 88.9 | 59.4 | 95.7 |
| | Adapter | 83M | 92.4 | 61.3 | 95.2 |
| | LoRA | 188.74M | 93.9 | 61.8 | 96.0 |
| | OFT | 39.32M | 86.4 | 44.2 | 43.7 |
| | IA3 | 0.492M | 90.3 | 57.9 | 93.8 |
| | MoReS-L | 0.328M | 89.8 | 57.7 | 93.5 |

Table 3: Results of multi-scale tasks.

### 5.4 MULTI-SCALE DATA FINE-TUNING

During visual instruction tuning, the scale of specific task datasets can vary significantly. To gain a comprehensive understanding of our method compared to other training approaches, we follow the methodology of (Chen et al., 2022) and randomly sample 1K, 5K, and 10K data points from each dataset, defining these as small-scale, medium-scale, and large-scale tasks, respectively. Given the limited resources available, we choose MoReS-L for fine-tuning.

Table 3 demonstrates that MoReS exhibits strong capabilities across all scales. Notably, in small-scale tasks, MoReS outperforms full fine-tuning performance while using only 575 times fewer parameters than LoRA and 8,475 fewer than full fine-tuning. In contrast, methods like OFT and IA3 fail to surpass full fine-tuning despite utilizing significantly more parameters. This result underscores the practicality of MoReS in real-world scenarios where data collection can be challenging, suggesting that MoReS is suitable for multi-scale visual instruction tuning.

### 5.5 ABLATION STUDIES

To gain deeper insights into our MoReS method, we conduct ablation studies focusing on its subspace choice and steered visual token ratio. We use LLaVA Steering-3B model as our baseline for comparison. Table 4 summarizes the results of two types of ablations.

First, concerning the choice of subspace rank, we found that a rank of 1 achieves the highest average performance of 81.8 across four visual tasks while also requiring the fewest parameters, specifically 0.164M. Second, regarding the steered visual token ratio, we varied this parameter from 100% (dense steering) to 1% (sparse steering). The results indicate that a ratio of 1% is optimal, yielding the best or near-optimal performance on four benchmarks while also significantly reducing inference overhead due to its sparse steering approach.

## 6 LLAVA STEERING FACTORY

We identified a pressing need for a comprehensive framework to systematically analyze and compare various model architectures and training strategies in MLLMs. The diversity of design choices and

| Subspace Rank | TP* | SciQA-IMG | VizWiz | IconQA-txt | IconQA-blank | Avg |
|---|---|---|---|---|---|---|
| 1 | 0.164M | 89.6 | 59.2 | 84.0 | 94.2 | 81.8 |
| 2 | 0.328M | 89.7 | 59.2 | 83.9 | 94.0 | 81.7 |
| 4 | 0.655M | 89.5 | 58.7 | 83.8 | 94.1 | 81.5 |
| 8 | 1.340M | 89.6 | 58.9 | 83.7 | 93.9 | 81.5 |

| Steered Visual Token Ratio | SciQA-IMG | VizWiz | IconQA-txt | IconQA-blank |
|---|---|---|---|---|
| 1% | 89.7 | 59.2 | 84.0 | 94.1 |
| 25% | 89.9 | 59.0 | 80.2 | 93.8 |
| 50% | 88.9 | 59.0 | 79.8 | 92.6 |
| 100% | 85.8 | 60.5 | 67.7 | 87.8 |

Table 4: Results of the subspace rank choice and steered visual token ratio. The grey shading indicates the best results and our selected parameters.

techniques complicates the standardization and understanding of how these components interact. Evaluating each method across different open-source models is often time-consuming and inconsistent due to implementation differences, which necessitate extensive data preprocessing and careful alignment between architectures and training recipes.

In the LLaVA Steering Factory, we establish standardized training and evaluation pipelines, along with flexible data preprocessing and model configurations. Our framework allows researchers to easily customize their models with various training strategies without the need for additional coding. We implement all mainstream LLMs and vision encoders, including multiple PEFT methods and our proposed MoReS technique. Furthermore, we support a wide range of benchmarks and integrate our intrinsic modality imbalance evaluation. The goal of the LLaVA Steering Factory is to facilitate research in MLLMs, particularly in addressing intrinsic modality imbalance to optimize visual instruction tuning.

An overview of the main components of the LLaVA Steering Factory is provided in Figure 5.

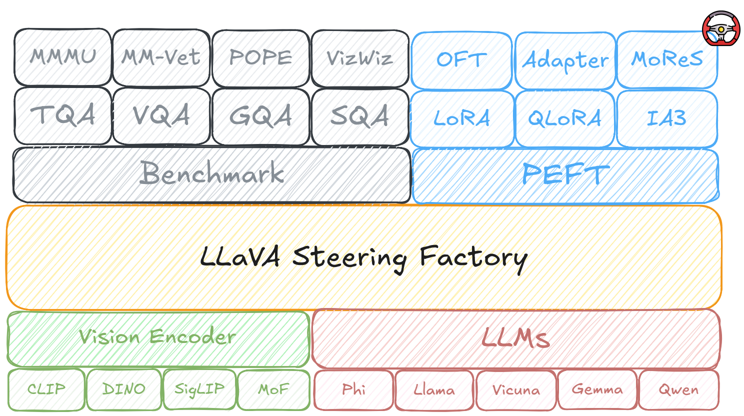


Figure 5: Architectural overview of the proposed LLaVA Steering Factory: A Modular Codebase for MLLMs.

## 7 CONCLUSION

This paper introduces Modality Linear Representation-Steering (**MoReS**), a novel method to significantly reduce the required number of trainable parameters during visual instruction tuning. The key idea behind MoReS is to re-balance visual and textual representations while still maintaining strong performance across a variety of downstream tasks. By integrating MoReS into LLaVA family models, comprehensive evaluation results confirm the effectiveness of the proposed solution. Hence, it further confirms our assertion that intrinsic modality rebalance would represent a promising new approach to optimizing visual instruction tuning.

To facilitate future research in the community, we also present the LLaVA Steering Factory, a versatile framework designed to enhance the development and evaluation of MLLMs with minimal coding effort. This framework enables customizable training configurations for various tasks and integrates analytical tools, such as attention score distribution analysis. This facilitates systematic comparisons among different methods and offers deeper insights into the intrinsic modality imbalance, ultimately contributing to more effective visual instruction tuning.

## References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL `https://arxiv.org/abs/2404.14219`.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf`.

Afra Amini, Tiago Pimentel, Clara Meister, and Ryan Cotterell. Naturalistic causal probing for morpho-syntax. *Transactions of the Association for Computational Linguistics*, 11:384–403, 2023.

Matan Avitan, Ryan Cotterell, Yoav Goldberg, and Shauli Ravfogel. Natural language counterfactuals through representation surgery, 2024. URL `https://arxiv.org/abs/2402.11355`.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. URL `https://www.adept.ai/blog/fuyu-8b`.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Guanzheng Chen, Fangyu Liu, Zaiqiao Meng, and Shangsong Liang. Revisiting parameter-efficient tuning: Are we really there yet? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2612–2626, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.168. URL https://aclanthology.org/2022.emnlp-main.168.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. *arXiv preprint arXiv:2403.06764*, 2024a.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

Yangyi Chen, Xingyao Wang, Hao Peng, and Heng Ji. A single transformer for scalable vision-language modeling. *arXiv preprint arXiv:2407.06438*, 2024b.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024c.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuoling Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms, 2024. URL https://arxiv.org/abs/2409.11402.

Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. *arXiv preprint arXiv:2406.11832*, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pp. 160–187. PMLR, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017a.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017b.

Clément Guerner, Anej Svete, Tianyu Liu, Alexander Warstadt, and Ryan Cotterell. A geometric notion of causal probing. *arXiv preprint arXiv:2307.15054*, 2023.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. Probing for the usage of grammatical number. *arXiv preprint arXiv:2204.08831*, 2022.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023. URL `https://arxiv.org/abs/2306.16527`.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023a.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36, 2024.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL `https://aclanthology.org/2021.acl-long.353`.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023b.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: **phi-1.5** technical report. *arXiv preprint arXiv:2309.05463*, 2023c.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf`.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, June 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. URL `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

Zikang Liu, Kun Zhou, Wayne Xin Zhao, Dawei Gao, Yaliang Li, and Ji-Rong Wen. Less is more: Data value estimation for visual instruction tuning. *arXiv preprint arXiv:2403.09559*, 2024c.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021.

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.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.

Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *Advances in Neural Information Processing Systems*, 36:79320–79362, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Erica K. Shimomoto, Edison Marrese-Taylor, Hiroya Takamura, Ichiro Kobayashi, and Yusuke Miyao. A subspace-based analysis of structured and unstructured representations in image-text retrieval. In Wenjuan Han, Zilong Zheng, Zhouhan Lin, Lifeng Jin, Yikang Shen, Yoon Kim, and Kewei Tu (eds.), *Proceedings of the Workshop on Unimodal and Multimodal Induction of Linguistic Structures (UM-IoS)*, pp. 29–44, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.umios-1.4. URL `https://aclanthology.org/2022.umios-1.4`.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioningwith reading comprehension. 2020.

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*, pp. 8317–8326, 2019.

Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roee Aharoni, Ryan Cotterell, and Ponnurangam Kumaraguru. Representation surgery: Theory and practice of affine steering, 2024. URL `https://arxiv.org/abs/2402.09631`.

Nishant Subramani, Nivedita Suresh, and Matthew E Peters. Extracting latent steering vectors from pretrained language models. *arXiv preprint arXiv:2205.05124*, 2022.

Qwen team. Qwen2-vl. 2024.

Yaqing Wang, Sahaj Agarwal, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan Awadallah, and Jianfeng Gao. AdaMix: Mixture-of-adaptations for parameter-efficient model tuning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5744–5760, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.388. URL https://aclanthology.org/2022.emnlp-main.388.

Muling Wu, Wenhao Liu, Xiaohua Wang, Tianlong Li, Changze Lv, Zixuan Ling, Jianhao Zhu, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. Advancing parameter efficiency in fine-tuning via representation editing. *arXiv preprint arXiv:2402.15179*, 2024a.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. Reft: Representation finetuning for language models, 2024b. URL https://arxiv.org/abs/2404.03592.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. Interpretability at scale: Identifying causal mechanisms in alpaca. *Advances in Neural Information Processing Systems*, 36, 2024c.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, Shrikant Kendre, Jieyu Zhang, Can Qin, Shu Zhang, Chia-Chih Chen, Ning Yu, Juntao Tan, Tulika Manoj Awalgaonkar, Shelby Heinecke, Huan Wang, Yejin Choi, Ludwig Schmidt, Zeyuan Chen, Silvio Savarese, Juan Carlos Niebles, Caiming Xiong, and Ran Xu. xgen-mm (blip-3): A family of open large multimodal models, 2024. URL https://arxiv.org/abs/2408.08872.

Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. Deco: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv preprint arXiv:2405.20985*, 2024.

Ting Yao, Yingwei Pan, Chong-Wah Ngo, Houqiang Li, and Tao Mei. Semi-supervised domain adaptation with subspace learning for visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

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. *arXiv preprint arXiv:2308.02490*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model, 2024. URL https://arxiv.org/abs/2401.02385.

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 A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023b.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. Tinyllava: A framework of small-scale large multimodal models, 2024a. URL https://arxiv.org/abs/2402.14289.

Xiongtao Zhou, Jie He, Yuhua Ke, Guangyao Zhu, Víctor Gutiérrez-Basulto, and Jeff Z Pan. An empirical study on parameter-efficient fine-tuning for multimodal large language models. *arXiv preprint arXiv:2406.05130*, 2024b.

Xingyu Zhu, Beier Zhu, Yi Tan, Shuo Wang, Yanbin Hao, and Hanwang Zhang. Selective vision-language subspace projection for few-shot clip. *arXiv preprint arXiv:2407.16977*, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

# A Appendix

## A.1 Theoretical Justification

Let $x_{\text{text}} \in \mathbb{R}^{d_t}$ be the text input embedding, $x_{\text{image}} \in \mathbb{R}^{d_v}$ be the visual input embedding, $R_{\text{text}} \in \mathbb{R}^D$ be the hidden representation for text, and $R_{\text{image}} \in \mathbb{R}^D$ be the hidden representation for the visual input. Define $W_q, W_k, W_v \in \mathbb{R}^{D\times D}$ as the query, key, and value projection matrices, and $W_o \in \mathbb{R}^{D\times D}$ as the output projection matrix. Let $A \in \mathbb{R}^{N\times N}$ represent the attention matrix, and $y \in \mathbb{R}^V$ be the output logits.

We present a theoretical analysis of the MoReS transformation and its effect on attention redistribution in multimodal models. The hidden representations for text and image inputs are computed as:

$$h_{\text{text}} = f_{\text{text}}(x_{\text{text}}), \quad h_{\text{image}} = f_{\text{image}}(x_{\text{image}}) \tag{9}$$

where $f_{\text{text}}$ and $f_{\text{image}}$ are encoding functions. The attention mechanism is characterized by scores:

$$A_{ij} = \text{softmax}\left(\frac{(h_i W_q)(h_j W_k)^T}{\sqrt{D}}\right) \tag{10}$$

with $W_q, W_k \in \mathbb{R}^{D\times D}$ being query and key projection matrices. Output generation follows:

$$y = W_o(C_{\text{text}} + C_{\text{image}}) \tag{11}$$

where $C_{\text{text}} = \sum_i A_{i,\text{text}}(h_i W_v)$ and $C_{\text{image}} = \sum_i A_{i,\text{image}}(h_i W_v)$.

The core of our approach is the MoReS transformation, defined as:

$$\text{MoReS}(h) = W_{\text{up}} \cdot \phi(h), \quad \text{where} \quad \phi(h) = \text{Linear}(h) - W_{\text{down}} h \tag{12}$$

Here, $W_{\text{up}} \in \mathbb{R}^{D\times d}$, $W_{\text{down}} \in \mathbb{R}^{d\times D}$, and $d < D$. When applied to the image representation, we obtain $h'_{\text{image}} = \text{MoReS}(h_{\text{image}}) + h_{\text{image}}$, leading to updated attention scores:

$$A'_{i,\text{image}} = \text{softmax}\left(\frac{(h_i W_q)(h'_{\text{image}} W_k)^T}{\sqrt{D}}\right) \tag{13}$$

This transformation is key to redistributing attention towards visual inputs. The effect of MoReS on the output can be quantified by examining the change magnitude:

$$\|\Delta y\|_2 = \|W_o(C'_{\text{image}} - C_{\text{image}})\|_2 \le \|W_o\|_2 \|C'_{\text{image}} - C_{\text{image}}\|_2 \tag{14}$$

where $C'_{\text{image}} = \sum_i A'_{i,\text{image}}(h'_{\text{image}} W_v)$. The significance of this change stems from the MoReS transformation's ability to amplify key visual features. Specifically, $\phi(h)$ extracts salient visual information in a subspace, which is then amplified by $W_{\text{up}}$ in the original space. This process ensures $\|h'_{\text{image}}\|_2 > \|h_{\text{image}}\|_2$, leading to increased $A'_{i,\text{image}}$ values for relevant visual features and larger magnitudes for $(h'_{\text{image}} W_v)$ terms in $C'_{\text{image}}$.

To ensure stability while allowing for this significant attention redistribution, we consider the Lipschitz continuity of the model:

$$\|f(h'_{\text{image}}) - f(h_{\text{image}})\|_2 \le L \|h'_{\text{image}} - h_{\text{image}}\|_2 \tag{15}$$

where L is the Lipschitz constant. This property bounds the change in the model's output, guaranteeing that the attention redistribution, while substantial, remains controlled and does not destabilize the overall model behavior.

A key advantage of the MoReS approach lies in its parameter efficiency. The transformation introduces $O(Dd)$ parameters, primarily from $W_{\text{up}}$, $W_{\text{down}}$, and the linear transformation in $\phi(h)$. This is significantly less than the $O(D^2)$ parameters required for fine-tuning all attention matrices in traditional approaches. The reduction in trainable parameters not only makes the optimization process more efficient but also mitigates the risk of overfitting, especially in scenarios with limited training data.

In conclusion, our theoretical analysis demonstrates that our MoReS effectively redistributes attention to visual inputs by operating in a carefully chosen subspace. This approach achieves a significant change in output generation while maintaining model stability and requiring fewer parameters than full fine-tuning, offering a balance between effectiveness and efficiency in enhancing visual understanding in MLLMs.

### A.2 Implementation Detail

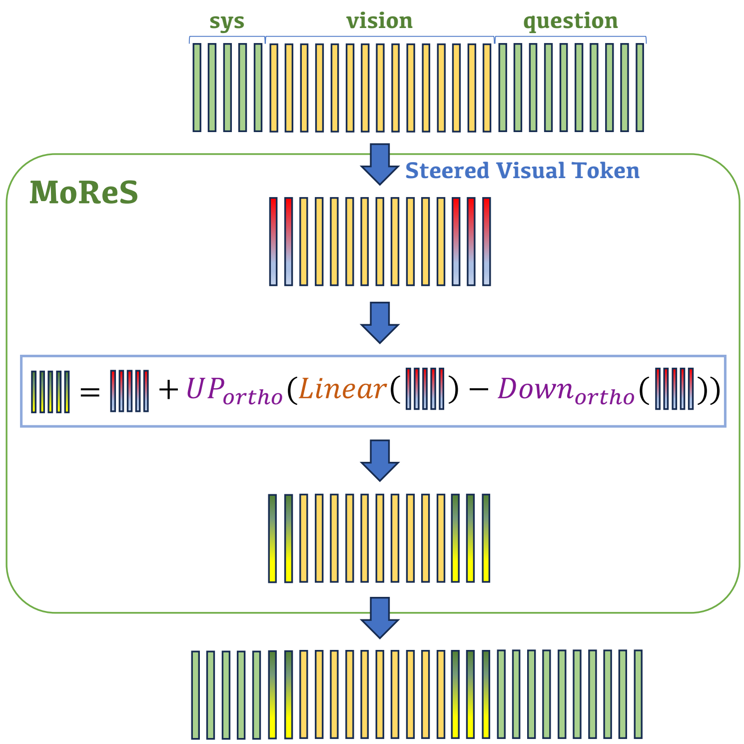


Figure 6: MoReS module flowchart.

Regarding the implementation, we have adopted a highly modular design for the LLM, integrating it with MoReS to enable precise steering at specific token locations. This modular approach ensures that the steering process operates with minimal computational overhead, making it both efficient and scalable. Additionally, the modular nature of this design allows for seamless integration with existing architectures and enables easy customization of steering strategies tailored to specific downstream tasks. To provide further clarity, we include a MoReS module flowchart (Figure 6) and an UML diagram (Figure 7) here, which detail the implementation process.

### A.3 Full Attention Maps

In this section, we provide the attention maps (Figure 8) during the decoding process across each layer. Notably, the distribution of visual attention remains sparse in these layers, with only a few tokens carrying the majority of the attention. This sparsity presents an opportunity for token pruning strategies, which can be leveraged to reduce inference overhead and improve computational efficiency. By selectively pruning tokens with lower attention scores, unnecessary computations can be

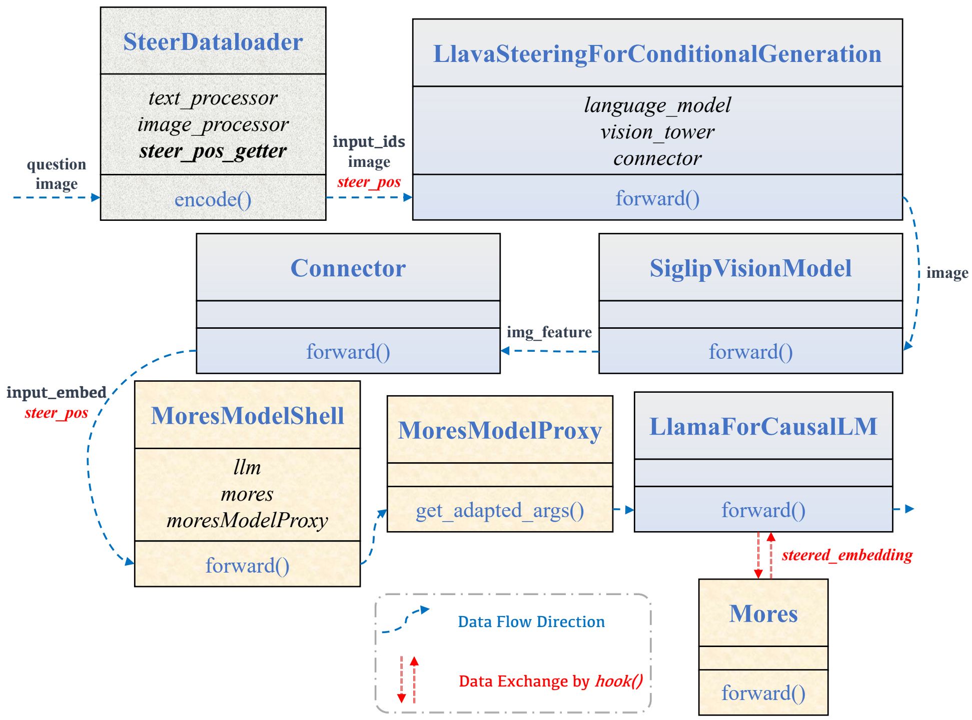


Figure 7: The UML diagram for MoReS

avoided, leading to faster and more efficient inference while maintaining the essential information needed for accurate predictions.

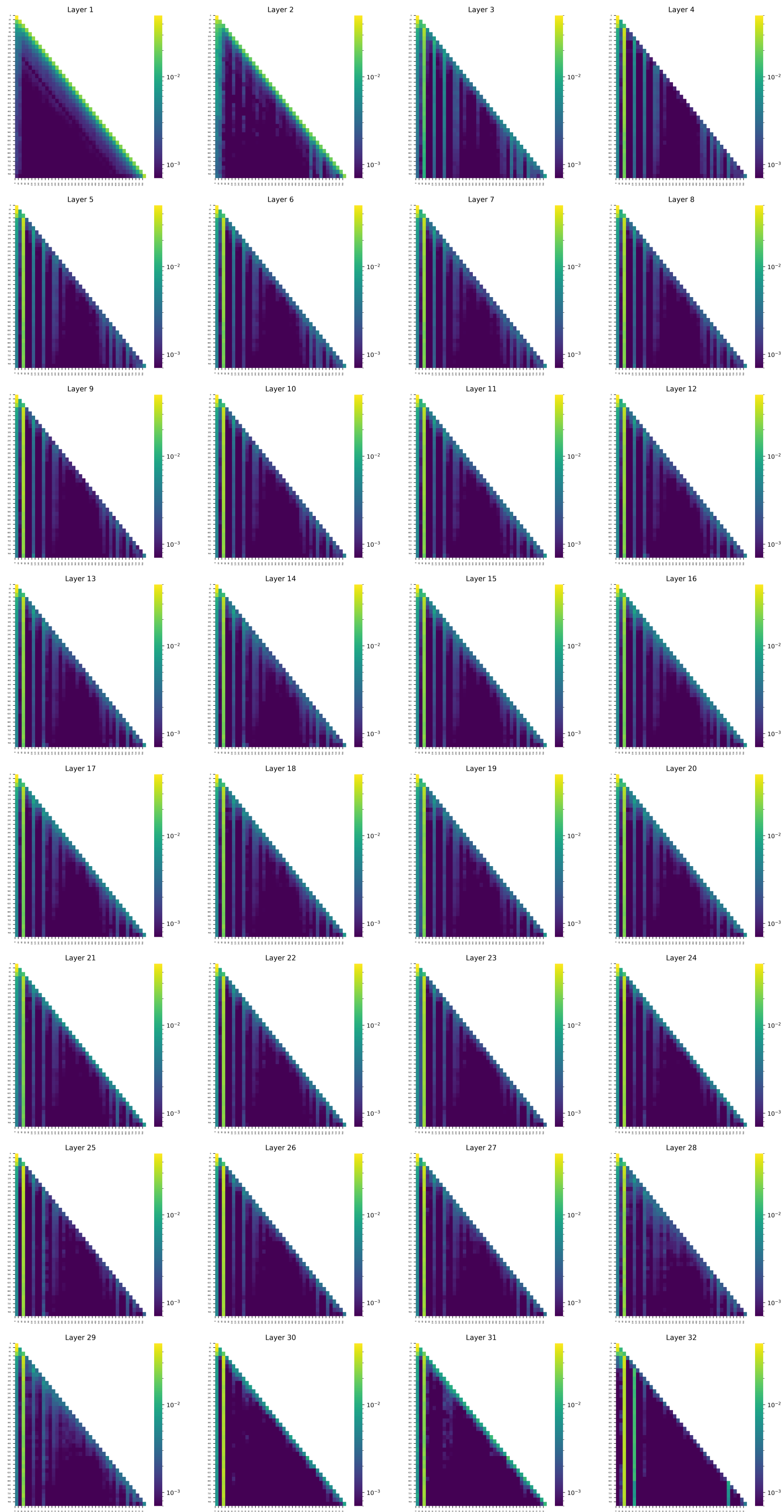


Figure 8: Full Attention Maps of Each Layer