MAMoE-LoRA: Modality-Aware Mixture of Experts Low-Rank Adaptation for QA task

Anonymous ACL submission

Abstract

Multimodal large language models (MLLMs) face challenges in efficiently adapting to diverse input types, such as text and images, due to the difficulty of processing heterogeneous modalities with a uniform approach. Traditional parameter-efficient fine-tuning (PEFT) methods, like LoRA, often treat all modalities equally, overlooking the need for modality-specific processing. To address this, we propose MAMoE-LoRA, a modality-aware framework that enhances expert specialization through a mixture-of-experts (MoE) architecture. Our approach organizes experts into three distinct pools: modality-specific experts for each input type, modality-shared experts for cross-modal integration, and always-active experts for consistent, domain-agnostic adaptation. We introduce an enhanced gating mechanism that utilizes causal-aware features and modality embeddings to intelligently route tokens to the most suitable experts. Additionally, we apply similarity regularization to maintain expert diversity and prevent overfitting. Experiments across multiple multimodal benchmarks demonstrate that MAMoE-LoRA achieves strong performance with minimal parameter overhead, requiring only 1.83–2.53% of trainable parameters while outperforming existing PEFT methods.

1 Introduction

The emergence of multimodal large language models (MLLMs) has revolutionized artificial intelligence, enabling unprecedented capabilities in vision-language understanding (Alayrac et al., 2022; Li et al., 2023). Models such as Qwen2.5-VL (Bai et al., 2025), GPT-4o (OpenAI et al., 2024), and LLaVA (Liu et al., 2023) demonstrate remarkable proficiency across diverse multimodal tasks. However, adapting these billion-parameter models to specialized domains presents significant computational challenges (Han et al., 2024).

Parameter-efficient fine-tuning (PEFT) methods address this challenge by adapting models while freezing most pre-trained parameters. Low-Rank Adaptation (LoRA) (Hu et al., 2022) has become particularly prominent, decomposing weight updates into low-rank matrices to reduce trainable parameters by orders of magnitude while achieving performance comparable to full fine-tuning. However, extending PEFT to multimodal scenarios introduces fundamental challenges. The heterogeneity of multimodal inputs—encompassing distinct feature distributions and processing requirements across text and vision—creates a complex optimization landscape that traditional PEFT approaches struggle to navigate.

Recent studies reveal that standard LoRA suffers from *modality competition* in multimodal settings (Wei et al., 2025; Huang et al., 2022): shared parameters optimized for one modality can degrade performance on another, and uniform parameter sharing fails to accommodate diverse vision-language task requirements (Shen et al., 2024). This highlights a critical tension between parameter efficiency and modality-specific adaptation. While unified sharing promotes efficiency, it constrains the model's ability to develop specialized representations for distinct modalities—particularly problematic for complex tasks requiring nuanced understanding of modality-specific patterns and cross-modal interactions (Hadji-Kyriacou and Arandjelovic, 2023).

We propose **Modality-Aware Mixture of Experts Low-Rank Adaptation (MAMoE-LoRA)**, a framework that reconciles parameter efficiency with modality-specific adaptation through hierarchical expert organization. Our approach maintains separate LoRA expert pools for each modality, complemented by shared experts for cross-modal knowledge transfer and always-active experts for universal representations (Liu et al., 2018). Three key innovations enable this: (1) *modality-specific*

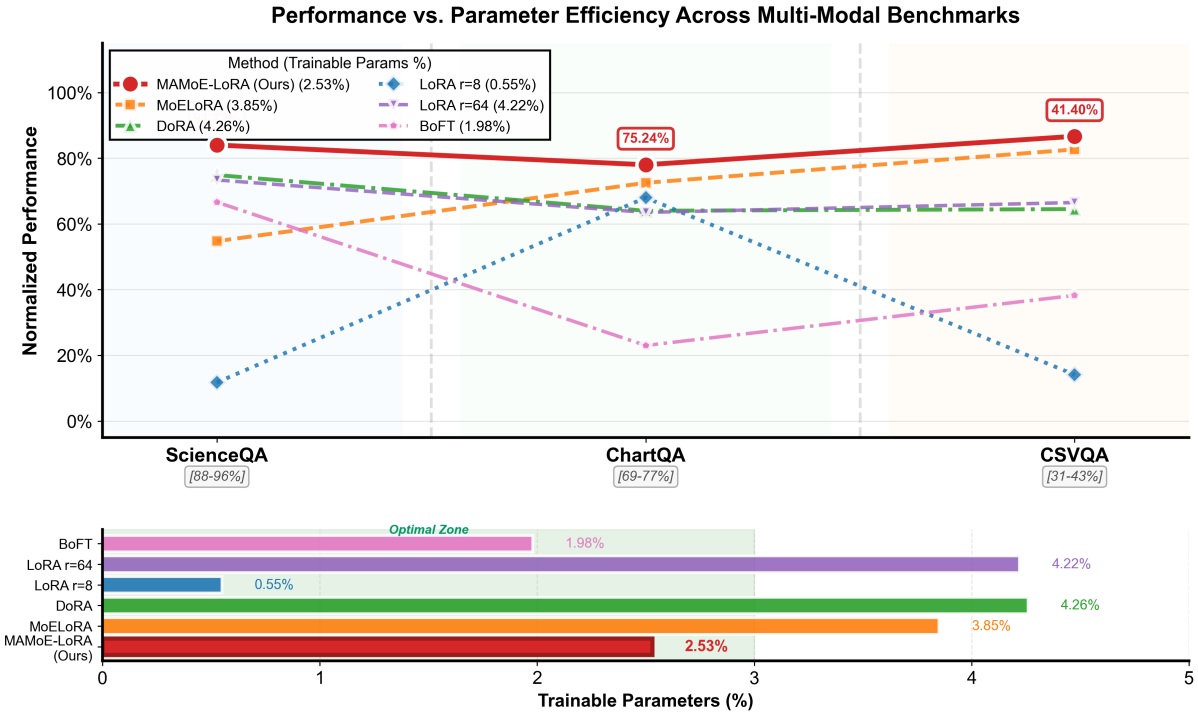

Figure 1: **Performance and parameter efficiency across multi-modal benchmarks.** *(Top)* Performance comparison on datasets with task-specific accuracy ranges. MAMoE-LoRA consistently achieves the best performance across all tasks. *(Bottom)* Trainable parameter overhead. MAMoE-LoRA attains optimal efficiency-performance trade-off at 2.53% parameters, outperforming methods with higher overhead (DoRA: 4.26%, LoRA r=64: 4.22%) while avoiding the performance degradation of extremely low-rank methods (LoRA r=8: 0.55%).

expert pools for specialized adaptation, (2) an *enhanced gating mechanism* incorporating causal-aware features for intelligent routing, and (3) *similarity regularization* enforcing expert diversity and preventing collapse.Our contributions are:

- A **modality-aware mixture-of-experts architecture** that organizes LoRA experts into modality-specific and shared pools, enabling specialized adaptation and effective cross-modal knowledge transfer.
- An **enhanced gating mechanism** that incorporates causal-aware feature extraction to improve expert routing under multimodal and temporal contexts.
- A **similarity regularization strategy** that explicitly enforces expert diversity across different modality-specific expert pools.
- Extensive experiments demonstrating **state-of-the-art performance** while updating only 1.83–2.53% of the model parameters.

As shown in Figure 1, MAMoE-LoRA consistently outperforms existing PEFT methods across diverse benchmarks while maintaining exceptional parameter efficiency, validating that explicit modality modeling is essential for effective multimodal adaptation. We release our code publicly [1].

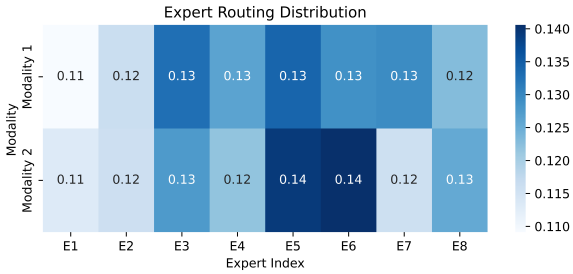

Figure 2: Expert routing distribution in traditional MoE-LoRA across two modalities. The uniform activation patterns indicate insufficient modality specialization.

2 Related Work

Parameter-Efficient Fine-Tuning (PEFT). Adapting large pre-trained models with minimal trainable parameters has become standard practice. Early approaches include adapter modules (Houlsby et al.,

[1] https://anonymous.4open.science/r/MAMoE-LoRA-1FC4

2019), which insert small trainable networks into transformer layers, and prompt-based methods like prefix-tuning (Li and Liang, 2021), which learn continuous input embeddings without modifying model weights. Most notably, LoRA (Hu et al., 2022) injects trainable low-rank matrices into each layer, reducing trainable parameters by orders of magnitude while achieving performance comparable to full fine-tuning. Recent extensions include UORA (Zhang et al., 2025), which introduces orthogonal initialization for improved stability, and L4Q (Jeon et al., 2025), which integrates quantization-aware training with LoRA for superior low-bit accuracy.

PEFT for Vision-Language Models. Extending PEFT to multimodal transformers presents unique challenges. In dual-encoder models like CLIP, adapter-based methods such as CLIP-Adapter, Tip-Adapter (Zhang et al., 2021), and XMAdapter (Zhou et al., 2022) fine-tune lightweight networks on modality-specific embeddings. However, these methods adapt each modality independently, "ignoring interactions between different modalities" (Seputis et al., 2024). To address this, Multi-Modal Adapter adds cross-attention for joint adaptation, while LLaMA-Adapter v2 (Gao et al., 2023) shows that LoRA-style modules can effectively align visual inputs to frozen LLMs. Recent work explores fine-grained alignment strategies: (Guo et al., 2025) achieve cross-modal semantic alignment with 0.25M parameters via prompt learning, while MoReS (Bi et al., 2025) steers visual representations through linear subspace transformations to rebalance text-vision attention.

Mixture-of-Experts for Adaptation. Combining parameter-efficient fine-tuning with Mixture-of-Experts (MoE) has proven effective for efficient model adaptation. Prior work integrates multiple LoRA experts into frozen LLMs with sparse routing to approach full fine-tuning performance using less than 1% trainable parameters (Zadouri et al., 2024).Subsequent methods explore LoRA-based MoE designs for multi-task learning (Li et al., 2024; Luo et al., 2024), as well as heterogeneous expert allocation and task-conditioned routing (Gao et al., 2025; Feng et al., 2024).

However, these MoE-based approaches are designed for unimodal scenarios and fail to address multimodal adaptation challenges. As shown in Figure 2, traditional MoE-LoRA exhibits three critical limitations: **(1) Lack of modality specialization**—expert activation weights show nearly identical patterns across modalities (0.11-0.14 range); **(2) Expert homogenization**—all experts exhibit similar frequencies ($\sim$0.125), suggesting functional equivalence; **(3) Insufficient modality discrimination**—maximum inter-modality difference is merely 0.03. These patterns reveal quasi-random expert selection in multimodal contexts, failing to exploit cross-modal heterogeneity. Our MAMoE-LoRA addresses these limitations through explicit modality-aware expert organization and enhanced routing mechanisms.

3 Methodology

We present Modality-Aware Mixture of Experts Low-Rank Adaptation (MAMoE-LoRA), a framework that efficiently adapts multimodal large language models through hierarchical expert organization and modality-aware routing.

3.1 Problem Formulation

Given a pre-trained vision-language model $\mathcal{M}$ with frozen parameters Θ, we adapt it for downstream multimodal tasks while maintaining parameter efficiency. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\} \in \mathbb{R}^{T\times d}$ denote the input sequence of length T with hidden dimension d, and $\mathbf{M} = \{m_1, m_2, \ldots, m_T\}$ represent modality indicators where $m_t \in \{0, 1, \ldots, K-1\}$ for K modalities.

The adaptation objective is:

$$\mathbf{Y} = \mathcal{F}(\mathbf{X}, \mathbf{M}; \Theta, \Delta\Theta), \tag{1}$$

where $\mathcal{F}$ is the adapted model and $\Delta\Theta$ represents trainable parameters enable modality-specific adaptation while preserving cross-modal interactions.

3.2 MAMoE-LoRA Architecture

Our framework comprises three key components: modality-enhanced gating, hierarchical expert organization, and causal-aware feature extraction. Figure 3 illustrates the overall architecture.

3.2.1 Modality-Enhanced Gating

Traditional MoE routing relies solely on input content. We incorporate modality information and temporal statistics to improve expert selection:

$$\text{Gate}(\mathbf{x}_t, \mathbf{s}_t) = \text{TopK}(\mathbf{W}_g(\mathbf{x}_t + \mathbf{e}_{m_t}) + \mathbf{b}_g + \alpha \cdot \mathbf{W}_s \mathbf{s}_t), \tag{2}$$

where $\mathbf{e}_{m_t} \in \mathbb{R}^d$ is the modality embedding for token t, $\mathbf{W}_g \in \mathbb{R}^{N\times d}$ and $\mathbf{W}_s \in \mathbb{R}^{N\times 4}$ are learnable weight matrices for N experts, and $\mathbf{s}_t \in \mathbb{R}^4$ captures distributional statistics.

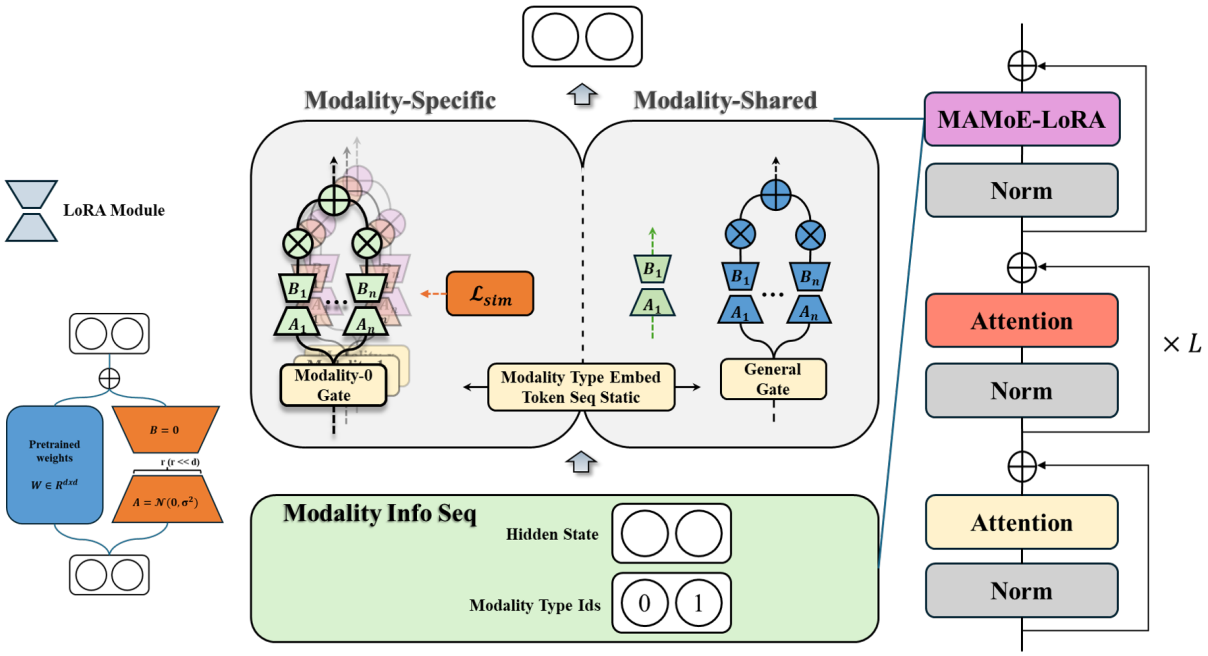

Figure 3: **MAMoE-LoRA architecture overview.** Modality embeddings and causal pooling statistics guide Top-K expert routing. Hierarchical expert organization includes modality-specific pools, shared pools, and always-active LoRA, integrated via residual connections.

Causal-Aware Feature Extraction To preserve autoregressive properties, we extract temporal context using causal pooling and statistical features capture distributional characteristics:

$$\hat{\mathbf{x}}_t = \frac{1}{t}\sum_{i=1}^{t}(\mathbf{x}_i + \mathbf{e}_{m_i}). \tag{3}$$

$$\mathbf{s}_t = [\mu_t, \sigma_t, \max_t, \min_t]^\top, \tag{4}$$

where μ_t, σ_t, $\max_t$, and $\min_t$ are mean, standard deviation, maximum, and minimum of $\hat{\mathbf{x}}_t$. This ensures routing decisions at position t depend only on current and previous tokens.

3.2.2 Hierarchical Expert Organization

We organize experts into three tiers to capture different levels of modality interactions:

Modality-Specific Experts For each modality $k \in \{0, 1, \ldots, K-1\}$, we maintain N LoRA experts $\{\mathcal{E}_1^{(k)}, \ldots, \mathcal{E}_N^{(k)}\}$:

$$\mathcal{E}_i^{(k)}(\mathbf{x}) = \mathbf{W}_{B,i}^{(k)}\mathbf{W}_{A,i}^{(k)}\mathbf{x} \cdot \frac{\alpha}{r}, \tag{5}$$

where $\mathbf{W}_{A,i}^{(k)} \in \mathbb{R}^{r\times d}$ and $\mathbf{W}_{B,i}^{(k)} \in \mathbb{R}^{d\times r}$ are down-projection and up-projection matrices with rank r and scaling factor α.

Modality-Shared Experts A shared expert pool $\{\mathcal{E}_1^{(s)}, \ldots, \mathcal{E}_N^{(s)}\}$ processes tokens from all modalities to capture cross-modal interactions:

$$w_{t,i}^{(s)}, \text{idx}_t^{(s)} = \text{Gate}^{(s)}(\mathbf{x}_t + \mathbf{e}_{m_t}, \mathbf{s}_t). \tag{6}$$

Always-Active Experts We include always-active shared experts $\mathcal{E}^{(a)}$ that contribute to all tokens, implemented as a single LoRA module.

3.2.3 Forward Computation

For token $\mathbf{x}_t$ with modality m_t, the final output combines all expert tiers:

$$\begin{aligned}\mathbf{y}_t = &\sum_{i\in\text{TopK}^{(m_t)}} w_{t,i}^{(m_t)}\mathcal{E}_i^{(m_t)}(\mathbf{x}_t) \\ &+ \sum_{j\in\text{TopK}^{(s)}} w_{t,j}^{(s)}\mathcal{E}_j^{(s)}(\mathbf{x}_t) + \mathcal{E}^{(a)}(\mathbf{x}_t),\end{aligned} \tag{7}$$

where $\text{TopK}^{(m_t)}$ and $\text{TopK}^{(s)}$ denote selected expert indices for modality-specific/shared pools.

3.3 Training Objective

The training objective combines task loss with similarity regularization:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda\mathcal{L}_{\text{sim}}, \tag{8}$$

where $\mathcal{L}_{\text{task}}$ is cross-entropy loss.

Table 1: Performance comparison on the ScienceQA benchmark. Accuracy (%) is reported for six categories: Natural Science (NAT), Social Science (SOC), Language Science (LAN), Text-only (TXT), Image-based (IMG), and No-context (NO). Results of the first four methods are taken from ScienceQA paper. Train.(%) denotes the ratio of trainable parameters. Bold and underlined numbers indicate the best and second-best results, respectively.

Method	**Train.(%)**	**NAT**	**SOC**	**LAN**	**TXT**	**IMG**	**NO**	**Average**
UnifiedQA BASE (CoT)	100	71.00	76.04	78.91	66.42	66.53	81.81	74.11
LLaMA-Adapter 6B	-	84.37	88.30	84.36	83.72	80.32	86.90	85.19
LLaVA 13B	100	90.36	95.95	88.00	89.49	88.00	90.66	90.92
GPT-4 w/ CoT	-	85.48	72.44	**90.27**	82.65	71.49	**92.89**	83.99
PEFT Methods on Qwen2.5-VL-3B								
Few-shot	0.00	57.15	49.04	67.27	54.25	51.66	68.99	60.12
LoRA r=8	0.55	90.00	92.80	83.64	89.54	87.90	86.06	88.94
LoRA r=64	4.22	94.36	<u>98.09</u>	89.45	94.13	94.25	91.15	93.87
LoHA	8.10	86.10	86.39	79.55	84.65	82.80	82.44	84.46
DoRA	4.26	<u>94.40</u>	97.98	89.91	<u>94.28</u>	<u>94.40</u>	91.29	<u>93.99</u>
BoFT	1.98	93.34	**98.31**	89.27	92.96	92.96	91.01	93.33
KRAdapter	1.06	85.08	83.46	78.18	83.08	81.45	80.97	82.95
MoELoRA	3.85	92.27	97.08	88.82	91.98	91.67	90.31	92.38
FusionLoRA	3.82	94.18	97.64	<u>90.18</u>	93.74	93.51	<u>91.85</u>	93.87
MAMoE-LoRA (Ours)	1.83	**95.83**	97.64	90.09	**95.65**	**95.54**	91.57	**94.72**

The similarity regularization encourages expert diversity across modalities:

$$\mathcal{L}_{\text{sim}} = \frac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{j \neq i} \text{CosSim}(\Theta^{(i)}, \Theta^{(j)}), \tag{9}$$

where $\Theta^{(i)}$ denotes concatenated parameters of all experts in modality i, preventing expert collapse while maintaining modality specialization.

3.4 Integration with Vision-Language Models

MAMoE-LoRA integrates into decoder layers as an additional adaptation branch. For each transformer layer l:

$$\mathbf{h}^{(l+1)} = \mathbf{h}^{(l)} + \text{MAMoE-LoRA}^{(l)}(\text{LN}(\mathbf{h}^{(l)})), \tag{10}$$

where $\mathbf{h}^{(l)}$ denotes hidden states at layer l and $\text{LN}(\cdot)$ is layer normalization. This residual integration leverages pre-trained representations while introducing minimal computational overhead.

4 Experiment

4.1 Experiment Setting

Datasets. To comprehensively evaluate the effectiveness of parameter-efficient fine-tuning methods on multi-modal large language models, we conduct experiments on three widely-adopted benchmarks that require joint understanding of visual and textual information: ScienceQA, which contains multiple-choice science questions spanning natural science, social science, and language science at elementary and high school levels (Saikh et al., 2022); ChartQA, a benchmark designed for question answering about charts requiring both visual perception and logical reasoning over visualized data (Masry et al., 2022); And CSVQA, a Chinese Multimodal Benchmark for Evaluating STEM Reasoning Capabilities of VLMs (Jian et al., 2025). These datasets collectively assess diverse capabilities of VLMs using PEFT method including scientific reasoning, chart comprehension, and structured data interpretation, providing a robust testbed for multi-modal understanding.

Baseline. We compare our proposed method against two categories of baselines. The first category consists of large-scale foundation models evaluated in a zero-shot setting without task-specific fine-tuning, including both open-source models (e.g., Qwen2.5-VL (Bai et al., 2025), LLaVA (Bi et al., 2025), GLM-4.6V (Team et al., 2025)) and closed-source models (e.g., GPT-4 (OpenAI et al., 2024), Gemini2.5 (Comanici et al., 2025)). These models represent the upper bound of pre-trained capabilities on multi-modal tasks. The second category comprises excellent parameter-efficient fine-tuning methods that adapt pre-trained models with minimal trainable parameters. This includes

Table 2: Comprehensive performance comparison on CSVQA dataset. The table includes excellent open-source and closed-source VLMs as baselines are taken from CSVQA papers, followed by PEFT methods applied to Qwen2.5-VL-3B. **Bold** indicates best performance among PEFT methods, underline indicates second-best.

Method	Model Size	Train. Size	Train. (%)	Overall	Bio	Chem	Math	Phys	Open	MC	Easy	Med	Hard
Open-source VLMs (Full Model)													
Deepseek-VL2	4.5B	-	-	7.00	6.20	7.60	4.50	8.00	6.00	-	-	-	-
LLaVA1.5-13B	13B	-	-	10.70	9.40	5.40	5.50	4.00	7.80	-	-	-	-
MonoInternVL	2B	-	-	7.30	9.10	9.20	10.90	3.00	9.80	-	-	-	-
Pixtral-12B	12B	-	-	15.30	8.80	8.60	10.00	5.00	10.90	-	-	-	-
Phi-4	14B	-	-	13.30	16.10	8.90	8.30	7.00	11.80	-	-	-	-
Gemma3-27B	27B	-	-	26.00	23.50	27.00	17.10	23.00	22.90	-	-	-	-
Internvl2-5-78B	78B	-	-	36.30	36.10	24.10	19.70	16.00	29.30	-	-	-	-
QVQ-72B	72B	-	-	40.70	41.30	33.70	32.00	32.00	36.90	-	-	-	-
Closed-source VLMs (Full Model)													
GPT-4o	-	-	-	23.60	28.00	23.50	23.50	20.60	18.00	24.00	-	-	-
Claude3.7	-	-	-	36.60	41.70	38.10	37.10	31.30	32.00	36.90	-	-	-
Gemini2.0-flash	-	-	-	44.10	45.00	45.50	47.60	39.80	46.00	44.00	-	-	-
o1	-	-	-	49.60	46.20	45.10	59.00	49.10	41.30	50.20	-	-	-
PEFT Methods on Qwen2.5-VL-3B													
Zero-shot	3.7B	-	-	0.73	3.16	0.00	0.00	0.00	0.00	0.78	1.11	0.72	0.00
LoRA (r=8)	3.7B	20M	0.55	32.69	48.42	54.46	13.13	17.80	0.00	34.97	42.22	34.41	2.27
LoRA (r=64)	3.9B	165M	4.22	38.98	50.53	57.43	25.25	25.42	7.41	41.19	47.78	41.22	6.82
KRAdapter	3.8B	20M	0.53	39.23	52.63	57.43	23.23	26.27	3.70	41.71	48.89	41.22	6.82
DoRA (r=64)	4.0B	166M	4.26	38.74	52.63	57.43	24.24	23.73	7.41	40.93	47.78	40.86	6.82
BoFT	3.8B	75M	1.98	35.59	46.32	51.49	24.24	22.88	0.00	38.08	46.67	36.56	6.82
FusionLoRA	4.0B	155M	3.82	40.92	54.74	58.42	26.26	27.12	3.70	43.52	47.78	43.73	9.09
MoELoRA	4.0B	156M	3.85	40.92	51.58	**59.41**	28.28	27.12	**7.41**	43.26	**50.00**	43.37	6.82
MAMoE-LoRA (Ours)	3.9B	99M	2.53	**41.40**	51.58	56.44	**34.34**	**26.27**	**7.41**	**43.78**	**50.00**	**44.09**	6.82

LoRA, which applies low-rank decomposition to weight updates (Hu et al., 2022); DoRA, which decomposes weight updates into magnitude and direction components (Liu et al., 2024a); LoHa, which employs Hadamard product-based low-rank adaptation for enhanced expressiveness (Hyeon-Woo et al., 2022); KRAdapter, which leverages the Khatri-Rao product to achieve higher effective rank in weight updates (Albert et al., 2025); BOFT, which uses butterfly-structured orthogonal transformations for parameter-efficient adaptation (Liu et al., 2024b); and MoELoRA, which considers LoRA as a Mixture of Experts architecture with dynamic expert selection (Li et al., 2024). These methods represent diverse design philosophies in PEFT, from low-rank approximation to orthogonal transformations and MoE architectures.

Implementation. We use a unified training framework with consistent hyperparameters across PEFT methods. For CSVQA, we use a 70/30 train-test split. Models are fine-tuned independently per dataset with performance averaged over three random seeds. For MoE-based methods, we set expert count $N \in \{2, 4, 8\}$ based on task complexity. Zero-shot baselines use task-specific prompts. Evaluation uses exact match accuracy for multiple-choice and relaxed accuracy for numerical answers. Training uses NVIDIA GPUs.

4.2 Result Analysis

Tables 1, 4, and 2 present comprehensive comparisons across all three benchmarks. Our analysis reveals several key findings regarding parameter efficiency, cross-task generalization, and architectural design choices. **Overall Performance.** MAMoE-LoRA achieves excellent results among PEFT methods: 94.72% on ScienceQA, 75.24% on ChartQA, and 41.40% on CSVQA, using only 1.83–2.53% trainable parameters. On ScienceQA, we surpass DoRA (93.99%) by 0.73 points with 2.43% fewer parameters. On ChartQA, we outperform MoELoRA (74.80%) by 0.44 points with 36.5% fewer parameters, validating our MoE architecture.

Comparison with Foundation Models. While closed-source models like Claude-3.7-Sonnet (82.31% on ChartQA) and o1 (49.60% on CSVQA) achieve higher absolute performance, MAMoE-LoRA demonstrates remarkable efficiency. On ScienceQA, our method achieves 94.72%, approaching GPT-4 w/ CoT (83.99%) and surpassing LLaMA-Adapter 6B (85.19%), while using only a

Table 3: Ablation study on ScienceQA. We report accuracy (%) and absolute drops (Δ) relative to the full model. Checkmarks indicate included components.

Method	Components					Accuracy (%)	
	Static Scale	Modal Pool	Sim Loss	Shared LoRA	General Pool	Img	Overall
MAMoE-LoRA (Full)	✓	✓	✓	✓	✓	**95.54**	**94.72**
w/o Static Scale		✓	✓	✓	✓	95.48 (-0.06)	94.69 (-0.04)
w/o Sim Penalty	✓	✓		✓	✓	95.04 (-0.50)	94.03 (-0.69)
w/o Shared LoRA	✓	✓	✓		✓	94.94 (-0.60)	93.63 (-1.09)
w/o General Pool	✓	✓	✓	✓		94.59 (-0.95)	93.37 (-1.35)
w/o Modal & Sim	✓			✓	✓	91.42 (-4.12)	91.86 (-2.86)

Table 4: Performance comparison on the ChartQA dataset. Results for both closed-source and open-source models are obtained from our own experiments, with detailed settings described in the appendix.

Method	Param (%)	Average
Closed-source Models		
Grok-3	–	48.02
GPT-4.1	–	73.16
Claude-3.7-Sonnet	–	82.31
Qwen-VL-Max	–	56.13
Open-source Models		
GLM-4.6v	–	38.02
Qwen2.5-VL-72B-Instruct	–	57.20
PEFT Methods (Base: Qwen2.5-VL-3B)		
LoRA ($r = 8, \alpha = 16$)	0.55	74.44
LoRA ($r = 64, \alpha = 128$)	4.22	74.08
LoHa ($r = 64$)	8.10	74.00
DoRA ($r = 64$)	4.26	74.12
FusionLoRA ($r = 256$)	3.82	72.92
BoFT	1.98	70.84
MoELoRA	3.85	<u>74.80</u>
KRAdapter	1.06	74.44
MAMoE-LoRA (Ours)	2.53	**75.24**

Table 5: Efficiency comparison on CSVQA. Training and inference time normalized to LoRA r=64.

Method	Params (%)	Acc. (%)	Train Time	Infer Time
LoRA r=64	4.22	38.98	1.00×	1.00×
KRAdapter	0.53	39.23	1.66×	1.49×
DoRA r=64	4.26	38.74	1.86×	1.96×
MoELoRA	3.85	40.92	1.08×	0.94×
MAMoE-LoRA	**2.53**	**41.40**	1.10×	0.95×

3.7B base model with 1.83% parameters fine-tuned. Against open-source models, MAMoE-LoRA substantially outperforms Qwen2.5-VL-72B-Instruct (57.20% on ChartQA) despite using 19× fewer parameters, validating that efficient fine-tuning compensates for model scale limitations.

Parameter Efficiency Analysis. The relationship between trainable parameters and performance reveals critical insights. On ScienceQA, MAMoE-LoRA (1.83%, 94.72%) outperforms LoRA r=64 (4.22%, 93.87%) with 57% fewer parameters, indicating superior parameter utilization. Conversely, extremely low-rank methods like LoRA r=8 (0.55%) achieve 88.94% on ScienceQA, substantially underperforming despite similar efficiency. This pattern holds across datasets: on ChartQA, LoRA r=8 achieves 74.44% while LoRA r=64 achieves 74.08%.

Domain-Specific Analysis. MAMoE-LoRA achieves top performance on most ScienceQA categories: 95.83% on NAT (vs. GPT-4's 85.48%), 95.65% on TXT, and 95.54% on IMG. BoFT slightly outperforms on SOC (98.31% vs. 97.64%), suggesting domain-specific architectural benefits. On CSVQA, MAMoE-LoRA excels in Mathematics (34.34%, +21.4% over MoELoRA's 28.28%) while maintaining competitive performance on Chemistry (56.44%) and Physics (26.27%), demonstrating superior quantitative reasoning.

Modality-Specific Performance. On ScienceQA, comparing IMG (95.54%), TXT (95.65%), and NO (91.57%) contexts reveals consistent high performance across modalities, with slightly lower scores on no-context questions that require pure knowledge recall. This pattern differs from baselines: GPT-4 shows larger variance (71.49% on IMG vs. 92.89% on NO), indicating less robust multi-modal integration. The consistent performance across modalities suggests that MAMoE-LoRA effectively captures cross-modal dependencies through expert specialization.

Key Takeaways. Our experimental results demonstrate that: (1) MAMoE-LoRA achieves optimal trade-offs between performance and parame-

ter efficiency across diverse multi-modal tasks; (2) Moderate parameter allocation (1.83-2.53%) with expert-based routing outperforms both extremely low-rank and high-rank monolithic adaptations.

4.3 Efficiency Analysis

Table 5 compares computational efficiency on CSVQA, with time normalized to LoRA r=64. MAMoE-LoRA achieves best accuracy (41.40%) using only 2.53% parameters with competitive efficiency (1.10× training, 0.95× inference). Despite dynamic expert routing, overhead remains comparable to MoELoRA while delivering +0.48% accuracy improvement. Extreme parameter reduction (KRAdapter: 0.53%) incurs substantial computational penalties (1.66×, 1.49×), while DoRA suffers the highest overhead (1.86×, 1.96×). MAMoE-LoRA thus achieves optimal balance for practical deployment.

4.4 Ablation Study

We systematically validate each component's contribution through ablation experiments on ScienceQA. Table 3 demonstrates that removing any component degrades performance, confirming the necessity of our hierarchical design.

Modality-Specific Pools. Removing component causes the largest performance drop (-2.86% overall, -4.12% on IMG), validating our core hypothesis that explicit modality specialization is essential. Without dedicated pools, shared experts face a representational bottleneck when handling divergent visual and textual reasoning requirements.

Always-Active LoRA. Removing component reduces performance by -1.35%, demonstrating its role in providing universal adaptation across all tokens and preventing routing failures.

Shared Expert Pool. Without shared experts, performance drops by -1.09%, confirming their importance for cross-modal knowledge transfer in questions requiring joint visual-textual reasoning.

Similarity Regularization. Removing the similarity loss causes -0.69% degradation. Without explicit diversity enforcement, experts converge to similar parameters through collapse, losing modality-specific specialization.

Causal Static Features. Removing static features results in -0.04% drop. While the impact is modest, these features enable context-dependent expert selection, particularly valuable when sequence context matters.

Synergistic Effects. The combined removal of modality pools and similarity loss (-2.86%) demonstrates these components work synergistically: modality specialization requires both architectural support (dedicated pools) and training objectives (diversity regularization).

5 Conclusion

We introduce MAMoE-LoRA, a modality-aware parameter-efficient fine-tuning method for multimodal large language models. MAMoE-LoRA organizes LoRA experts into hierarchical pools—modality-specific experts for specialized adaptation, shared experts for cross-modal integration, and always-active experts for universal representations. This organization enables element-wise modality processing while maintaining efficient parameter usage. The enhanced gating mechanism incorporates modality embeddings and causal-aware features to route tokens intelligently, while similarity regularization prevents expert collapse and maintains functional diversity.

We demonstrate the efficiency of MAMoE-LoRA, which achieves strong performance with substantially reduced trainable parameters compared to traditional PEFT methods. Moreover, since the hierarchical expert structure is designed to explicitly model modality differences, it maintains inference efficiency while avoiding the representational bottlenecks of uniform adaptation approaches. The effectiveness of MAMoE-LoRA as a multimodal adaptation framework is further supported by experimental results across diverse benchmarks. MAMoE-LoRA consistently achieves superior performance in vision-language tasks, demonstrating enhanced adaptability compared to existing LoRA variants and MoE-based methods that lack modality-aware design.

Limitations

Our work has several limitations that suggest directions for future research.

Model Scale. Experiments are conducted on models up to 3.7B parameters (Qwen2.5-VL-3B). Scalability to larger models (e.g., 70B+ parameters) and their expert utilization patterns remain unexplored.

Modality and Task Coverage. Our evaluation focuses on two-modality vision-language tasks; extension to richer modalities (audio, video, 3D) and broader task types (generation, dense prediction) remains future work.

Expert Architecture. Our design uses uniform expert allocation across modalities. Heterogeneous configurations—varying expert counts, ranks, or adaptive Top-K routing based on modality characteristics—may yield further improvements.

Interpretability. While our method improves accuracy, the internal cross-modality mechanisms and the specialization of modality-specific experts remain underexplored.

Future work should investigate scaling to larger models, extending to multi-modality scenarios beyond vision-language, designing adaptive expert allocation strategies, and conducting interpretability studies to understand emergent specialization patterns.

References

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, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Paul Albert, Frederic Z. Zhang, Hemanth Saratchandran, Anton van den Hengel, and Ehsan Abbasnejad. 2025. Towards higher effective rank in parameter-efficient fine-tuning using khatri-rao product. *CoRR*, abs/2508.00230.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923.

Jinhe Bi, Yujun Wang, Haokun Chen, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. 2025. LLaVA steering: Visual instruction tuning with 500x fewer parameters through modality linear representation-steering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15230–15250, Vienna, Austria. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Wenfeng Feng, Chuzhan Hao, Yuewei Zhang, Yu Han, and Hao Wang. 2024. Mixture-of-loras: An efficient multitask tuning method for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 11371–11380. ELRA and ICCL.

Chongyang Gao, Kezhen Chen, Jinmeng Rao, Ruibo Liu, Baochen Sun, Yawen Zhang, Daiyi Peng, Xiaoyuan Guo, and V. S. Subrahmanian. 2025. Mola: Moe lora with layer-wise expert allocation. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 5097–5112. Association for Computational Linguistics.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. Llama-adapter V2: parameter-efficient visual instruction model. *CoRR*, abs/2304.15010.

Yongbin Guo, Shuzhen Li, Zhulin Liu, Tong Zhang, and C.L.Philip Chen. 2025. A parameter-efficient and fine-grained prompt learning for vision-language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31346–31359, Vienna, Austria. Association for Computational Linguistics.

Avelina Asada Hadji-Kyriacou and Ognjen Arandjelovic. 2023. Context-peft: Efficient multi-modal, multi-task fine-tuning. *CoRR*, abs/2312.08900.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Trans. Mach. Learn. Res.*, 2024.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. 2022. Modality competition: What makes joint training of multi-modal network fail in deep learning? (provably). In *International*

Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, volume 162 of *Proceedings of Machine Learning Research*, pages 9226–9259. PMLR.

Nam Hyeon-Woo, Moon Ye-Bin, and Tae-Hyun Oh. 2022. Fedpara: Low-rank hadamard product for communication-efficient federated learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Hyesung Jeon, Yulhwa Kim, and Jae-Joon Kim. 2025. L4Q: parameter efficient quantization-aware fine-tuning on large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 2002–2024. Association for Computational Linguistics.

Ai Jian, Weijie Qiu, Xiaokun Wang, Peiyu Wang, Yunzhuo Hao, Jiangbo Pei, Yichen Wei, Yi Peng, and Xuchen Song. 2025. CSVQA: A chinese multimodal benchmark for evaluating STEM reasoning capabilities of vlms. *CoRR*, abs/2505.24120.

Dengchun Li, Yingzi Ma, Naizheng Wang, Zhiyuan Cheng, Lei Duan, Jie Zuo, Cal Yang, and Mingjie Tang. 2024. Mixlora: Enhancing large language models fine-tuning with lora based mixture of experts. *CoRR*, abs/2404.15159.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4582–4597. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024a. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Weiyang Liu, Zeju Qiu, Yao Feng, Yuliang Xiu, Yuxuan Xue, Longhui Yu, Haiwen Feng, Zhen Liu, Juyeon Heo, Songyou Peng, Yandong Wen, Michael J. Black, Adrian Weller, and Bernhard Schölkopf. 2024b. Parameter-efficient orthogonal finetuning via butterfly factorization. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2247–2256. Association for Computational Linguistics.

Tongxu Luo, Jiahe Lei, Fangyu Lei, Weihao Liu, Shizhu He, Jun Zhao, and Kang Liu. 2024. Moelora: Contrastive learning guided mixture of experts on parameter-efficient fine-tuning for large language models. *CoRR*, abs/2402.12851.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: a novel resource for question answering on scholarly articles. *Int. J. Digit. Libr.*, 23(3):289–301.

Dominykas Seputis, Serghei Mihailov, Soham Chatterjee, and Zehao Xiao. 2024. Multi-modal adapter for vision-language models. *CoRR*, abs/2409.02958.

Ying Shen, Zhiyang Xu, Qifan Wang, Yu Cheng, Wenpeng Yin, and Lifu Huang. 2024. Multimodal instruction tuning with conditional mixture of lora. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 637–648. Association for Computational Linguistics.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 69 others. 2025. Glm-4.5v and glm-4.1v-thinking: Towards versatile

multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Yake Wei, Yu Miao, Dongzhan Zhou, and Di Hu. 2025. Moka: Multimodal low-rank adaptation for mllms. *CoRR*, abs/2506.05191.

Ted Zadouri, Ahmet Üstün, Arash Ahmadian, Beyza Ermis, Acyr Locatelli, and Sara Hooker. 2024. Pushing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2021. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *CoRR*, abs/2111.03930.

Xueyan Zhang, Jinman Zhao, Zhifei Yang, Yibo Zhong, Shuhao Guan, Linbo Cao, and Yining Wang. 2025. UORA: uniform orthogonal reinitialization adaptation in parameter efficient fine-tuning of large models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 11709–11728. Association for Computational Linguistics.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16795–16804. IEEE.

A Prompting and LLM-based Inference Details

This appendix describes the prompting strategy and large language model (LLM)-based inference and evaluation pipeline used in our experiments, particularly for ChartQA-style multimodal question answering. All prompts are fixed across experiments to ensure reproducibility and fair comparison.

A.1 System Prompts

We employ predefined system prompts to guide LLM behavior during inference. These prompts are not modified during training or evaluation.

A.1.1 English Science Question Answering Prompt

The following system prompt is used for science-related multiple-choice question answering tasks, where each question may optionally include an image (e.g., charts or diagrams):

```
You are a highly intelligent assistant specialized in answering science-related multiple-choice questions. Each question may include an image.
Your task is to analyze the question, consider the provided image if available, and give the answer.
If an explanation is required, provide a concise and clear explanation for your choice.
Format your response as follows:
Answer: <Your Answer>
Explanation: <Your Explanation (if applicable)>.
Always ensure your answers are based on scientific knowledge and logical reasoning.
(safe and brief mode)
```

This prompt enforces a structured output format and encourages concise, scientifically grounded reasoning.

A.2 Evaluation Prompt

To automatically assess the correctness of model-generated answers, we employ a separate LLM as an evaluator with a fixed evaluation prompt:

```
You are an AI assistant tasked with evaluating the correctness of answers generated by another AI model. You will be provided with the correct answer and the model's response.
If the response is correct, return 'Correct'.
If the response is incorrect, return 'Incorrect'.
Format your response as follows: Result: <Correct/Incorrect> Explanation: <Your Explanation (if applicable)>
```

The evaluator strictly performs answer matching and does not generate task outputs.

A.3 LLM-based Inference and Evaluation Pipeline

For ChartQA inference, we adopt a two-stage pipeline consisting of LLM-based reasoning and LLM-based answer verification.

A.3.1 Inference Stage

Given an input sample consisting of a question and an optional chart image, we construct the inference input as follows:

- **Text input**: the question text.
- **Image input**: the associated chart image, if available.
- **System prompt**: the science QA prompt described in Section A.1.

The LLM generates a structured response containing an answer and, when applicable, a brief explanation.

A.3.2 Evaluation Stage

The generated response is subsequently evaluated by a separate LLM using the evaluation prompt described in Section A.2. Given the ground-truth answer and the model response, the evaluator outputs a binary judgment:

$$\text{Result} \in \{\text{Correct}, \text{Incorrect}\}. \tag{11}$$

This automated evaluation protocol reduces ambiguity in answer matching and ensures consistent evaluation across different models.

—

A.4 Models and Decoding Settings

Unless otherwise specified, we use the following models and decoding configurations:

- **Evaluation model**: GPT-4.1-mini
- **Maximum tokens**: 64 for inference, 16 for evaluation
- **Decoding strategy**: greedy decoding
- **Streaming**: disabled

—

A.5 Reproducibility Statement

All prompts are fixed and shared across models. No prompt tuning, prompt optimization, or prompt-based learning is applied in any experiment. Therefore, all reported performance gains are attributed solely to the proposed **MAMoE-LoRA** adaptation framework rather than prompt engineering.

A.6 Generation Parameters

Unless otherwise specified, we use the following decoding parameters for all LLMs:

- Temperature: 0.7
- Top-p: 0.9
- Max tokens: 64

Table 6: Experiment settings.

Hardware/Software	**Setting**
OS	Ubuntu 20.04.1 LTS
CPU	Intel Core i9-10900K
GPU	RTX 3090 $\times$ 2
Python	3.10.0
PyTorch	2.5.1
Global Batch size	128
claude-3-7-sonnet(API)	claude-3-7-sonnet-20250219
LLM response evaluator	gpt-4.1-mini
Optimizer	AdamW
Epoch	3
Loss function	Cross Entropy Loss
Learning rate	3×10^{-4}
Learning rate schedule	Cosine
BF16	True
weight decay	0.01
warmup ratio	0.03
gradient accumulation steps	4
gradient checkpointing	True
image min pixels(qwen-vl)	256 * 28 * 28
image max pixels(qwen-vl)	1024 * 28 * 28
tf32	True
Flash Attention 2	True
number of routed expert	(2, 4, 8)
number of activate expert	(1, 2)