Multimodal Instruction Tuning with Conditional Mixture of LoRA

Ying Shen[♠] Zhiyang Xu[♠]
Qifan Wang[♢] Yu Cheng[♣] Wenpeng Yin[♡] Lifu Huang[♠]

[♠]Virginia Tech [♢]Meta AI [♣]The Chinese University of Hong Kong [♡]The Pennsylvania State University

Motivation

- **Parameter-Efficient Fine-tuning (PEFT):** Enabling efficient adaptation of pre-trained large models to various downstream applications by only fine-tuning a small fraction of model's parameters
- **PEFT in Multimodal Instruction Tuning:** Fine-tuning *a limited portion* of shared parameters for diverse multimodal instruction tasks simultaneously is likely to lead to task interference

> Task Interference: Sharing all parameters among different tasks results in performance degradation for a subset of tasks.

Our research seeks to explore and address task interference in parameter-efficient multimodal instruction tuning. Specifically, we aim to answer two critical research questions:

(1) Does task interference exist in parameter-efficient multimodal instruction tuning?

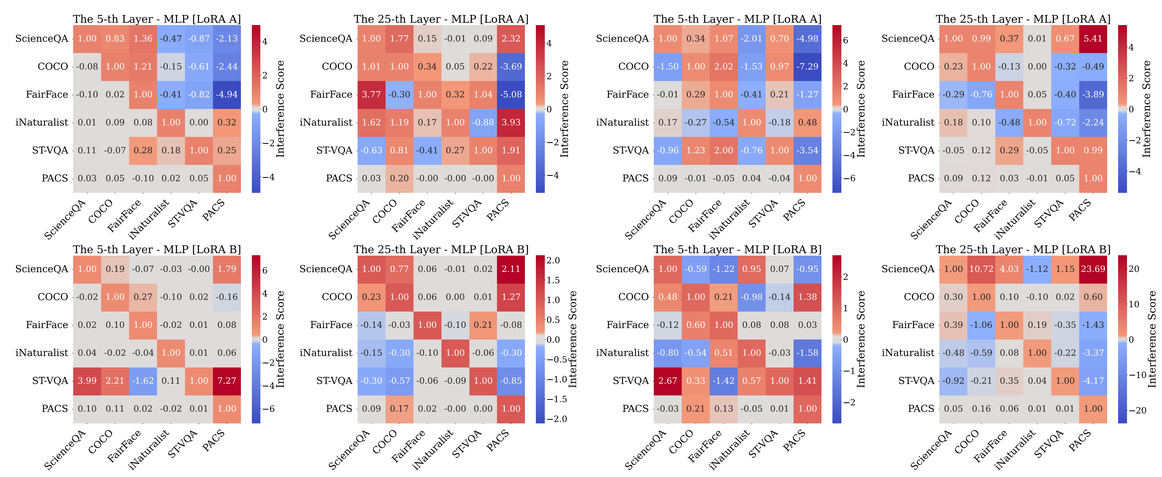
(2) How can we effectively mitigate this issue for robust performance across various tasks?

Task Interference in Multimodal Instruction Tuning with LoRA

We investigate task interference in parameter-efficient multimodal instruction tuning by analyzing gradient direction conflicts between task pairs. We compute the task interference matrix $\mathcal{I} \in \mathbb{R}^{M \times M}$ for LoRA decomposition matrices A and B, where M is the number of tasks and $\mathcal{I}_{i,j}$ quantify the interference of task j on task i.

Our results demonstrate significant task interference in both shallow and deep Transformer layers for LoRA A and B, with negative influences (blue) suggesting that learning one task can hinder another. Conversely, positive effects (red) indicate that one task's learning may enhance another's performance.

These findings highlight notable task interference in parameter-efficient multimodal instruction tuning and reinforce the need for effective adaption methods to ensure robust performance across diverse multimodal tasks.



(a) MLP (b) Self-Attention

Figure 1. **The Task Interference Score $\mathcal{I}$ for LoRA decomposition matrices A and B.** Each cell in the heatmap corresponds to the average interference score $\mathcal{I}_{i,j}$ of task j (column) on the task i (row). A blue hue indicates a negative impact of task j on task i, whereas a red hue signifies a positive impact.

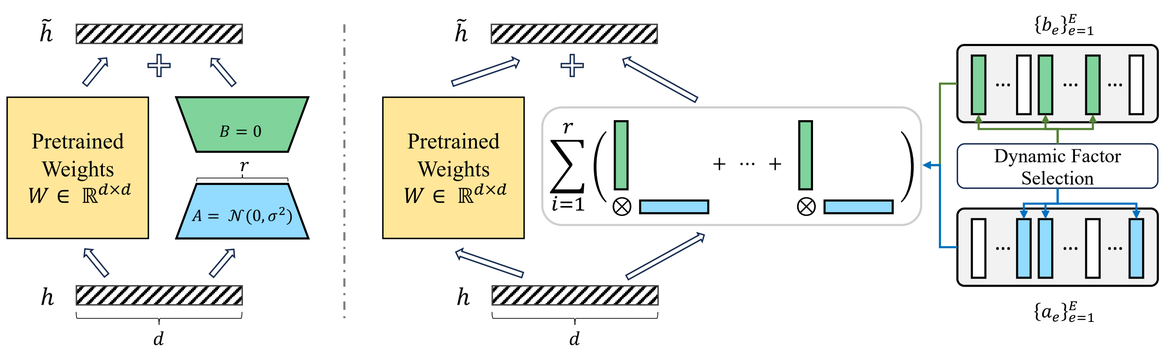
Conditional Mixture-of-LoRA (MixLoRA)

We propose **Conditional Mixture-of-LoRA (MixLoRA)** which leverages low-rank decomposition factors as dynamically chosen experts to construct tailored decomposition matrices A and B for specific input instances. We can represent the weight adjustment matrices ΔW from LoRA via tensor decomposition:

$$\Delta W = BA = \sum_{i=1}^{r} b_i \otimes a_i, \tag{1}$$

where $\{a_i, b_i\}_{i=1}^{r}, a_i \in \mathbb{R}^{d_{in} \times 1}, b_i \in \mathbb{R}^{d_{out} \times 1}$ are the rank r decomposition factors of $\Delta W \in \mathbb{R}^{d_{out} \times d_{in}}$.

Leveraging the concept that ΔW can be expressed as the sum of outer products of low-rank decomposition factors a_i and b_i, MixLoRA introduces a **Dynamic Factor Selection** module. This module dynamically constructs unique ΔW for specific inputs by selecting r appropriate factors from an expanded pool of decomposition factors $\{a_e\}_{e=1}^{E}, \{b_e\}_{e=1}^{E}, E > r$.



(a) LoRA (b) MixLoRA

MixLoRA dynamically constructs the LoRA A and B through two main components.

First, two **Independent Factor Selection (IFS)** routers, independently select r relevant factors to form adaptation matrices LoRA A and B, ensuring precise, instance-specific adaptations.

Second, a **Conditional Factor Selection (CFS)** router further refines LoRA B's selection by conditioning the selection for B also on the factors chosen for LoRA A, promoting a coherent adaptation process.

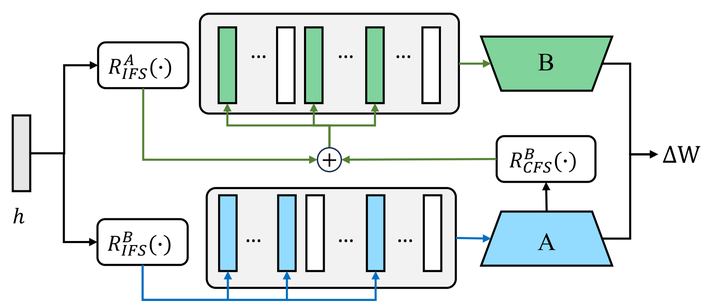


Figure 2. Dynamic Factor Selection in MixLoRA.

Results and Discussion

Model	Factors	Rank	MME	Text-VQA	VSR	SNLI-VE	CIFAR-10	CIFAR-100	MNIST	Pope	MMAvg
$\text{LLaVA}_{\text{Align}}$	-	-	1110.82	32.62	50.16	34.51	80.00	58.04	52.79	59.10	52.46
LLaVA_{FT}	-	-	1587.26	37.26	53.76	43.35	92.97	63.73	94.27	80.82	66.59
LoRA	-	2	1291.20	**39.86**	51.88	31.80	85.51	**49.23**	79.22	76.72	59.17
LoRA	-	4	1345.86	39.44	53.19	33.08	86.62	47.36	80.89	**76.89**	59.64
LoRA	-	8	1312.87	39.20	53.27	36.36	88.92	46.88	82.95	75.48	60.44
LoRA	-	16	1381.23	39.22	**53.60**	36.11	87.31	45.60	**85.92**	75.16	60.42
LoRA	-	32	**1393.67**	39.20	52.95	**44.56**	**90.10**	45.90	83.42	72.33	**61.21**
MixLoRA	16	2	1417.83	39.82	52.13	**35.38**	90.14	**58.05**	**85.98**	73.86	62.19
MixLoRA	32	2	**1459.15**	**40.46**	**52.62**	35.04	**91.02**	57.95	85.26	**78.31**	**62.95**
MixLoRA	16	4	1443.82	**40.66**	**52.70**	**43.10**	**91.59**	57.28	85.25	78.13	**64.10**
MixLoRA	32	4	**1509.61**	40.42	49.18	36.69	91.40	**59.27**	**87.68**	**78.48**	63.30
MixLoRA	16	8	1485.26	39.92	**52.70**	**40.74**	**92.85**	53.96	82.95	75.31	62.63
MixLoRA	32	8	**1485.48**	**40.02**	51.15	37.77	91.12	**60.25**	**86.64**	**78.87**	**63.69**

Table 1. **Zero-shot Multi-modal Evaluation.** $\text{LLaVA}_{\text{Align}}$ indicates the stage-one LLaVA-v1 with only feature alignment but not visual instruction tuning, and LLaVA_{FT} is the fully fine-tuned LLaVA using the same Vision-Flan dataset. The **MMAvg** column denotes the average performance across seven multimodal datasets, except for MME.

Comparison with LoRA

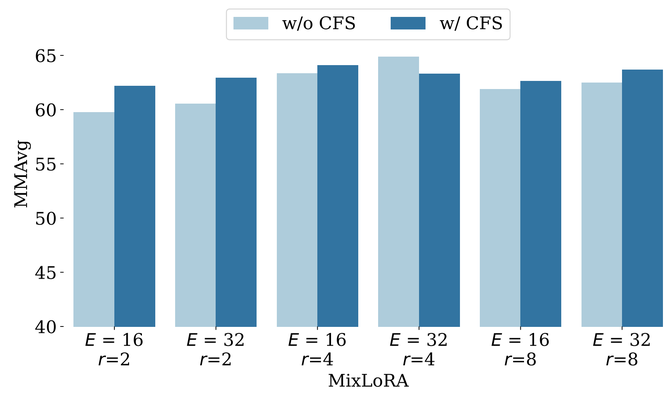
MixLoRA consistently outperforms LoRA at the same ranks on both MME and other multi-modal tasks and even surpasses LoRA at higher ranks.

Increase the Number of Rank / Factors

MixLoRA exhibits notable performance improvements as the rank number increased from 2 to 4 with a fixed factor number. When the rank number is constant, there is a general trend of enhanced performance for MixLoRA.

Effect of Conditional Factor Selection

Incorporating the Conditional Factor Selection router in general consistently improves the performance across different factor and rank settings.

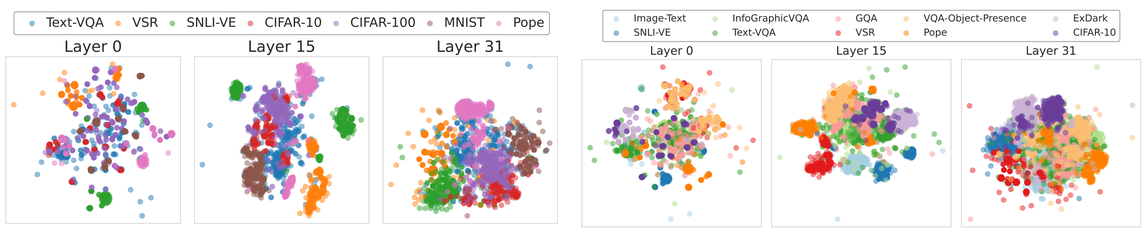


T-SNE Visualization of Factor Selection

We visualize the factor selection patterns in MixLoRA (E = 32, r = 8):

(a) Instances from identical tasks tend to cluster, indicating the effectiveness of an instance-based routing strategy in assigning diverse factor sets across tasks.

(b) MixLoRA effectively activates factors analogous to those employed in similar training tasks, suggesting that the model can adapt its factor selection strategies to unseen tasks based on its training on similar seen tasks.



(a) T-SNE Visualization of Factor Selection Distribution.

(b) T-SNE Visualization of Factor Selection for Seen and Unseen Tasks. Seen tasks in the first row are color-matched with their unseen counterparts in the second row.

Mitigation of Task Interference

Compared to $\text{LoRA}_{\text{Specialist}}$, conventional LoRA exhibits varying degrees of performance. In contrast, MixLoRA suffers less from performance degradation and demonstrates more consistent and robust performance across different tasks, suggesting its effectiveness in reducing task interference.

Model	Factors	Rank	ScienceQA	COCO	FairFace	iNaturalist	ST-VQA	PACS	AVG
$\text{LoRA}_{\text{Specialist}}$	-	4	64.33	77.67	54.67	58.67	44.67	99.00	66.50
$\text{LoRA}_{\text{Specialist}}$	-	16	67.33	76.33	59.00	60.00	46.33	99.00	68.00
LoRA	-	4	57.67	76.33	59.67	57.00	42.33	99.33	65.39
LoRA	-	16	59.67	73.00	59.33	58.33	43.67	99.00	65.50
MixLoRA	16	4	60.67	78.67	59.00	61.00	44.33	99.33	67.17

Table 2. **Multi-modal Evaluation on Seen Tasks.** $\text{LoRA}_{\text{Specialist}}$ represents the specialist LoRA model fine-tuned for each seen task individually. The **AVG** column denotes the average performance across six seen tasks.

Paper

Please check out our paper for more details!

Zhiyang Xu and Ying Shen contributed equally to this work.

MultiInstruct