SiGMA: Sign-Guided Merging and Adaptation for Multimodal Continual Instruction Tuning

Keonhee Park and Gunhee Kim

Seoul National University
keonhee.park@vision.snu.ac.kr, gunhee@snu.ac.kr

Abstract. Multimodal Continual Instruction Tuning (MCIT) is crucial for adapting Multimodal Large Language Models (MLLMs) to evolving a sequence of downstream tasks. Prior methods mostly utilize Mixture-of-Experts or expansion–merge approach, primarily focusing on catastrophic forgetting, yet they still suffer from negative interference during inference, where newly learned updates overwrite useful prior knowledge and degrade overall performance. To address this, we propose **SiGMA** (**Si**gn-**G**uided **M**erging and **A**daptation), a simple yet effective framework that mitigates negative interference with two components: sign-guided adaptive tuning during training and sign-guided merging at inference. Sign-guided adaptive tuning reduces collisions with past knowledge and learns the current task with minimal drift, mitigating severe forgetting. Sign-guided merging further improves consolidation by selectively scaling salient parameters to preserve and amplify useful task-specific knowledge. Experiments on UCIT and DCL benchmarks show that SiGMA significantly reduces negative interference and outperforms state-of-the-art MCIT methods. Our code is available at SiGMA.

Keywords: Continual Learning · Transfer Learning · Multi-modality

1 Introduction

Recently, Multimodal Large Language Models (MLLMs) [1,6,23,24] have gained significant attention for their strong instruction-following ability and effective vision–language alignment, achieving impressive performance across diverse tasks [21,26,33]. MLLMs are typically trained in pre-training followed by instruction tuning; pre-training learns vision–text aligned representations, while instruction tuning enables the model to follow user instructions for response generation. This training allows for zero-shot capability on unseen instructions. However, such zero-shot ability often fails to meet domain-specific needs, making additional instruction tuning necessary to adapt MLLMs to downstream tasks.

Multimodal Continual Instruction Tuning (MCIT) [5, 8, 15, 35, 38] enables MLLMs to continually acquire novel knowledge while preserving previously learned knowledge. One key well-known challenge in MCIT is catastrophic forgetting [28], where learning new tasks significantly degrades performance on previously learned tasks. In the context of MLLMs, catastrophic forgetting can manifest as the loss

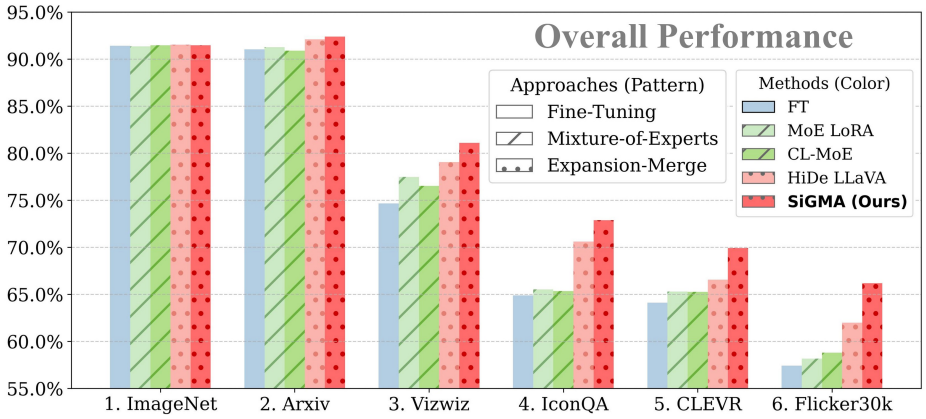

Fig. 1: Comparison of overall performance on UCIT benchmark. We report a holistic metric over the continual instruction tuning process, computed as the average accuracy on all previously seen tasks after training each task. While existing MoE and expansion–merge approaches gradually suffer from performance degradation due to negative interference and severe forgetting, SiGMA, which explicitly addresses negative interference, consistently maintains high overall performance.

of previously learned representation knowledge and degradation of instruction-following capability. As a result, the model may generate responses with incorrect content and fail to comply with user-specified constraints.

Conventional MCIT methods typically rely on LoRA [14] to adapt MLLMs to new tasks while preserving previously learned knowledge. Most prior work mitigates catastrophic forgetting through either Mixture-of-Experts (MoE) [5, 15] or expansion–merge strategies [8, 35]. MoE distributes learned knowledge across experts to reduce forgetting, whereas expansion–merge strategies train a task-specific LoRA for each task, freeze it after training, and merge the learned LoRAs at inference to consolidate task-specific knowledge. As shown in Figure 1, despite their effectiveness in mitigating forgetting, both approaches remain vulnerable to negative interference [2], where newly learned knowledge may overwrite earlier knowledge and degrade comprehensive performance throughout the continual learning process. In consequence, beyond preventing catastrophic forgetting, mitigating negative interference is essential for robust MCIT.

To this end, we propose the SiGMA (**Si**gn-**G**uided **M**erging and **A**daptation) framework that combines two simple yet effective strategies: *sign-guided adaptive tuning* during training and *sign-guided merging* at inference. The core idea of SiGMA is to decouple LoRA weights into general (task-invariant) and specific (task-variant) subspaces based on the parameter-wise sign pattern. From our empirical observation, parameters that share the same sign across tasks can be treated as a general subspace, whereas those with differing signs can be treated as a specific subspace. During training, sign-guided adaptive tuning reduces collisions with previously learned knowledge and adapts the current LoRA weights

by leveraging both the pre-trained MLLM and the historical general subspace, encouraging task learning with minimal drift. In inference, sign-guided merging integrates multiple LoRA weights by selecting salient parameters from the specific subspace and scaling them to strengthen useful task-specific knowledge.

Through this dual strategy, SiGMA can mitigate performance degradation throughout the continual instruction tuning process by explicitly addressing negative interference. Since negative interference occurs in consolidating previously learned weights, merely preserving prior knowledge intact cannot tackle this interference. As shown in Figure 1, the proposed SiGMA achieves promising improvements in mitigating performance degradation due to the negative interference while effectively preserving previously acquired representation knowledge.

Finally, our contributions are summarized as follows:

- We propose SiGMA, a simple yet effective sign-decoupling framework that aims to mitigate negative interference in continual instruction tuning. To the best of our knowledge, it is the first work to introduce the sign-based weight decoupling idea in continual instruction tuning.
- During training, the sign-guided adaptive tuning reduces knowledge collisions by adapting current LoRA weights to the general subspace, acquiring novel representations with minimal drift from prior knowledge.
- At inference, the sign-guided merging selectively amplifies crucial parameters in the specific subspace to retain its useful knowledge while reducing the impact of negative interference.
- SiGMA achieves state-of-the-art results on the UCIT (Unseen Continual Instruction Tuning) [8] and DCL (Domain Continual Learning) [39] benchmarks, and we further validate its effectiveness through extensive ablation studies and in-depth analysis experiments.

2 Related Work

Multimodal Large Language Models. Instruction tuning has been crucial in aligning large language models [17, 34] with human intention across diverse contexts. Recent visual instruction tuning [23,24] has extended this capability to multimodal large language models (MLLMs) [3, 7, 40], by integrating visual and text inputs for vision-language reasoning. However, most MLLMs are trained in an offline and static manner. This prevents MLLMs from adapting to the non-stationary nature of real-world tasks and evolving user requirements. In this work, we study multimodal continual instruction tuning, aiming to continuously improve the instruction following ability of MLLMs without severe forgetting that arises in sequential adaptation.

Multimodal Continual Instruction Tuning. Multimodal Continual Instruction Tuning (MCIT) requires MLLMs to continually follow instructions across successive tasks while preserving previously learned instruction-following abilities. Prior MCIT methods primarily rely on Mixture-of-Experts [16, 31] or expansion–merge strategies to adapt to new tasks while preserving earlier ones:

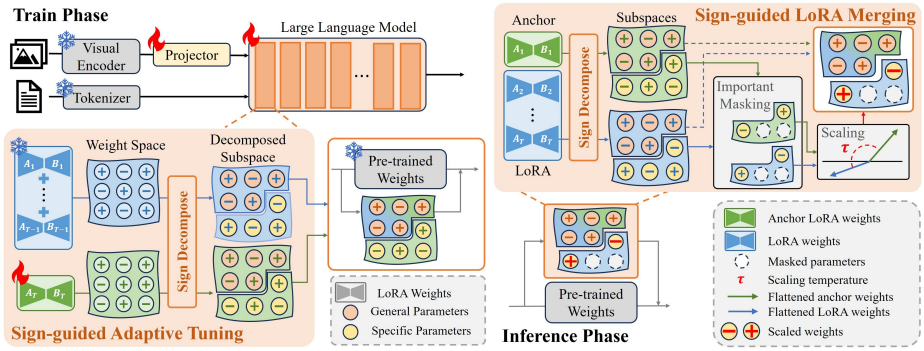

Fig. 2: Overview of SiGMA. During training, sign-guided adaptive tuning (left) decomposes prior weights into general and task-specific subspaces to convey transferable knowledge. During inference, sign-guided LoRA merging (right) consolidates weights by selecting and scaling salient task-specific parameters, preserving crucial knowledge.

CoIN [5] introduces an MCIT benchmark of eight datasets and proposes MoE-LoRA to mitigate catastrophic forgetting. CL-MoE [15] employs a dual-momentum MoE to route global and local experts. O-LoRA [35] leverages orthogonal subspace learning with orthogonal loss. HiDE-LLaVA [8] adopts task-specific expansion with task-general fusion.

While existing approaches can mitigate severe forgetting, they overlook negative interference during inference-time consolidation, where newly learned updates can overwrite useful prior knowledge. The proposed SiGMA aims to alleviate the negative interference via sign-guided merging at inference and mitigate severe parameter drift through sign-guided adaptive tuning during training.

3 Methodology: SiGMA

We first formulate the multimodal continual instruction tuning (MCIT) problem and present the base LoRA pipeline (§ 3.1). We then analyze negative interference, which can be observed in the form of sign-conflicted parameters (§ 3.2). Motivated by this, we introduce sign-based weight decoupling, the core mechanism of SiGMA (§ 3.3). Building on this, we detail sign-guided adaptive tuning, which leverages previously learned weights to efficiently adapt the model to the current task (§ 3.4). Finally, we present our sign-guided LoRA merging strategy, which selectively masks and amplifies important parameters to retain task-specific knowledge while minimizing negative interference (§ 3.5). Figure 2 shows an overall framework of the proposed SiGMA.

3.1 Preliminary

MCIT aims to learn from a continuous data stream while mitigating catastrophic forgetting. We consider a pre-trained MLLM parameterized by θ to be learned

from a sequence of T tasks. Each task $t \in \{1, \ldots, T\}$ is defined by a dataset $D_t = \{(x_{vis}^{t,i}, x_{ins}^{t,i}, x_{ans}^{t,i})\}_{i=1}^{N_t}$, where N_t denotes the number of image-text paired samples and x_{vis}, x_{ins}, and x_{ans} denote the input image, instruction, and answer tokens, respectively. Given an image–text input pair with an target answer of length L, the model is optimized to predict the next token in an autoregressive manner using the standard objective function:

$$\mathcal{L}^t = -\sum_{l=1}^{L} \log p(x_{ans}^l \mid x_{vis}^t, x_{ins}^t, x_{ans}^{<l}; \theta). \tag{1}$$

When learning task t, the goal of MCIT is to acquire novel knowledge from $\mathcal{D}_t$ while minimizing the forgetting of previous knowledge from $\mathcal{D}_{1:t-1}$. However, naive fine-tuning suffers from severe catastrophic forgetting, which disrupts not only the knowledge acquired from prior tasks but also the pre-trained foundational capabilities. Therefore, it is crucial to retain this foundational knowledge with the task-specific knowledge acquired during MCIT.

To achieve this, we utilize Parameter-Efficient Fine-Tuning (PEFT), which adapts large models to downstream tasks with minimal parameters. In the context of MCIT, Low-Rank Adaptation (LoRA) [14] is widely adopted due to its effectiveness. Given a frozen pre-trained weight matrix $W \in \mathbb{R}^{d_{out} \times d_{in}}$, LoRA models the learnable weight $\Delta W = B \times A$, using two low-rank matrices $A \in \mathbb{R}^{r \times d_{in}}$ and $B \in \mathbb{R}^{d_{out} \times r}$, where the low-rank $r \ll \min(d_{out}, d_{in})$. The modified weight matrix W' during training is formulated as

$$W' = W + \Delta W = W + BA. \tag{2}$$

By optimizing only A and B, LoRA achieves efficient adaptation while keeping W frozen. In this work, we adopt an expansion-merge strategy as a base model. We train task-specific LoRA modules (A_t, B_t) for each task t. At inference, we consolidate these parameters into unified weights $A' = \sum_{i=1}^{t} A_i$ and $B' = \sum_{i=1}^{t} B_i$ to enable comprehensive inference across all trained tasks.

3.2 Negative Interference in Continual Instruction Tuning

Existing MCIT methods widely employ Mixture-of-Experts (MoE) or expansion-merge strategies to mitigate catastrophic forgetting. While effective in preserving learned knowledge, both approaches suffer from negative interference, where newly acquired knowledge conflicts with prior one. The negative interference harms the model's knowledge representation, as the weight consolidation can overwrite and distort preserved knowledge. Therefore, addressing negative interference is as crucial as mitigating catastrophic forgetting in MCIT.

Our research starts from the observation that the negative interference is highly correlated with the parameter signs. Given previously learned weights W_{old} and newly learned weights W_{new}, we partition the parameter space into a sign-aligned subspace R_{alig} and a sign-conflicted subspace R_{conf}. When consolidating weights, we assume that the effects differ significantly across subspaces.

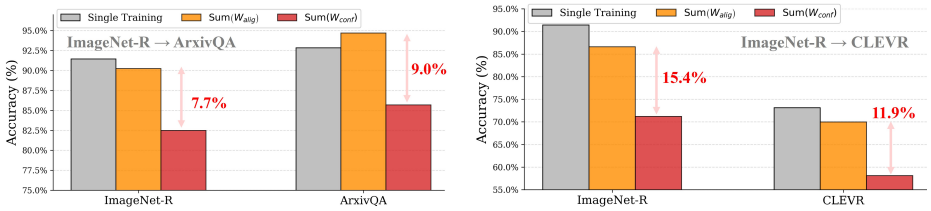

Fig. 3: Impact of sign-based subspaces on continual learning performance. We train separate LoRA modules for ImageNet-R, ArxivQA, and CLEVR to preserve task-specific knowledge, and then merge them at inference. W_{alig} represents merging only sign-aligned weights, while W_{conf} represents merging only sign-conflicted weights.

In the sign-aligned subspace, weights share the same sign, so their magnitudes tend to accumulate, potentially reinforcing the learned knowledge. In contrast, in the sign-conflicted subspace, weights have opposing signs, which can cause cancellation during merging, reducing the effective magnitude and degrading the corresponding representation knowledge. This is formulated as follows:

$$R_{alig} = \{i|\text{sgn}(W_{old}^{(i)}) = \text{sgn}(W_{new}^{(i)})\},\ R_{conf} = \{j|\text{sgn}(W_{old}^{(i)}) \neq \text{sgn}(W_{new}^{(j)})\}, \quad (3)$$

$$|W_{old}^{\mathcal{R}_{alig}} + W_{new}^{\mathcal{R}_{alig}}| = |W_{old}^{\mathcal{R}_{alig}}| + |W_{new}^{\mathcal{R}_{alig}}| \quad \text{(Positive Interference)}, \quad (4)$$

$$|W_{old}^{\mathcal{R}_{conf}} + W_{new}^{\mathcal{R}_{conf}}| \ll |W_{old}^{\mathcal{R}_{conf}}| + |W_{new}^{\mathcal{R}_{conf}}| \quad \text{(Negative Interference)}. \quad (5)$$

The negative interference due to opposing signs leads to a diminishing magnitude of the consolidated weights. This cancellation during weight consolidation may cause severe performance degradation despite the intention of consolidation to harmonize prior knowledge.

For empirical validation, we conduct experiments to analyze the effects of sign-aligned and sign-conflicted subspaces. We employ "Single Training" (*i.e.*, independent training on each task) as an upper-bound baseline, representing the optimal performance for a specific task. As shown in Figure 3, merging the sign-aligned subspace W_{alig} preserves performance close to the baseline and, for ArxivQA, even surpasses the baseline, showing the aligned subspace tends to contribute to positive interference. In contrast, merging the sign-conflicted subspace W_{conf} causes severe degradation. This hints that opposing signs tend to cause negative interference, where weight cancellation may lead imparing of prior knowledge and pose the weights in a suboptimal state, compromising both old task (*e.g.*, ImageNet-R) and new tasks (*e.g.*, ArxivQA and CLEVR). We provide diverse case studies in the supplementary material that further demonstrate the observed behaviors of sign-aligned and sign-conflicted subspaces.

3.3 Sign-based Weight Decoupling

To address the negative interference in MCIT based on the observation in § 3.2, we propose sign-based weight decoupling, which partitions weights into sign-aligned and sign-conflicted subspaces. Using the proposed decomposition, we

not only leverage the sign-aligned subspace during the training to transfer prior positive knowledge, but also handle the sign-conflicted subspace in inference to mitigate negative interference. To the best of our knowledge, this is the first sign-based decoupling approach explicitly designed to mitigate negative interference.

Given the weights W_{old} trained up to the previous task and the weights W_{cur} for the current task, we compute binary general mask M^g for the aligned subspace and the specific mask M^s for the conflicting subspace as follows:

$$M^g = \mathbb{I}\left(\operatorname{sgn}\left(W_{old}\right) = \operatorname{sgn}\left(W_{new}\right)\right), \quad M^s = \mathbb{I}\left(\operatorname{sgn}\left(W_{old}\right) \neq \operatorname{sgn}\left(W_{new}\right)\right), \quad (6)$$

where $\mathbb{I}(\cdot)$ and $\operatorname{sgn}(\cdot)$ denote indicator function and sign function, respectively. We decompose W_{old} into a general subspace $W^g_{old} = M^g \odot W_{old}$ and a specific subspace $W^s_{old} = M^s \odot W_{old}$, where $\odot$ denotes the element-wise product. The M^g and M^s masks form a partition over the parameters (*i.e.*, $M^g \odot M^s = 0$ and $M^g + M^s = \mathbf{1}$), ensuring that the sum of the general and specific subspaces exactly reconstructs the original old weights W_{old}.

Intuitively, W^g_{old} contains the subset of parameters where the signs of W_{old} and W_{new} are aligned. Since this alignment indicates that the optimization landscapes of the previous and current tasks are consistent, W^g_{old} represents task-invariant knowledge that can serve as a stable basis for learning the new task. In contrast, W^s_{old} consists of parameters with opposing signs, indicating task-specific knowledge that is crucial for the previous tasks but conflicts with the information of current task.

3.4 Sign-Guided Adapative Tuning

Building upon the sign-based weight decoupling strategy, we propose sign-guided adaptive tuning. This approach facilitates the learning of the current task t by effectively leveraging transferable knowledge from previous tasks while filtering out conflicting information.

During the training of task t, we first construct a consolidated representaion of prior knowledge. We aggregate the low-rank matrices A and B separately across all $t-1$ previous tasks:

$$\Delta W_{prev} = \left(\sum_{i=1}^{t-1} B_i\right)\left(\sum_{i=1}^{t-1} A_i\right) = \sum_{i=1}^{t-1} B_i A_i + \sum_{i \neq j} B_i A_j. \quad (7)$$

By combining weights from different tasks, we explore a broader latent subspace that captures transferability across tasks, rather than limiting the model to the disjoint union of previous task spaces (*i.e.*, applying only first term in Eq. 7). Then, we dynamically decompose ΔW_{prev} based on the evolving current LoRA weights, $\Delta W_{cur} = B_t A_t$, to build a general subspace ΔW^g_{prev}. We strictly utilize the sign-aligned general subspace to transfer knowledge to the current task. The modified forwarding process during training is as follows:

$$W' = W_\theta + \Delta W^g_{prev} + \Delta W_{cur}, \text{ where} \quad (8)$$

$$\Delta W^g_{prev} = M^g \odot \Delta W_{prev}, \quad M^g = \mathbb{I}\left(\operatorname{sgn}\left(\Delta W_{prev}\right) = \operatorname{sgn}\left(\Delta W_{cur}\right)\right). \quad (9)$$

W_θ denotes the frozen pre-trained weights. The proposed sign-guided adaptive tuning has the property of dynamic adaptation. Although the aggregated prior weight ΔW_{prev} remains static, the extracted general subspace ΔW^g_{prev} updates throughout the training process. As the current weight ΔW_{cur} updates, the sign alignment shifts, allowing the model to adaptively retrieve different parameters for the prior knowledge that align with the current optimization direction.

In summary, the proposed sign-guided adaptive tuning provides two key benefits. First, it facilitates the exploration of new knowledge by using the transferable parameters in ΔW^g_{prev} as a basis. Second, it imposes a soft regularization constraint. By ensuring alignment with the general subspace of prior knowledge, new task parameters do not deviate drastically from the previously learned weight space. Consequently, this alignment minimizes the conflict between ΔW_{cur} and ΔW_{prev}, reducing the risk of negative interference during final consolidation.

3.5 Sign-Guided LoRA Merging

While the proposed sign-guided adaptive tuning effectively aligns transferable knowledge during training, the remained task-specific subspaces tend to suffer from negative interference. Since these subspaces contain unique, task-dependent knowledge defined by divergent optimization directions, simply aggregating them leads to severe negative interference, where conflicting parameters disrupt each other. To address this at the inference stage, we propose sign-guided LoRA merging, which selectively strengthens salient task-specific parameters while filtering out noise-like parameters, ensuring that the unique characteristics of each task are preserved within the unified model.

After training up to task t, weight merging is needed to infer the model to represent the overall learned knowledge. To partition between general and specific subspaces consistently across all tasks, we use the LoRA weights from the first task ΔW_1 as the anchor weights. ΔW_1 should provide the stable basis for defining the parameter space, since it serves as the adaptation of the pre-trained model and is utilized through all subsequent training tasks in adaptive tuning. For every subsequent task $k \in \{2, ..., t\}$, we decouple the weights ΔW_k based on the anchor ΔW_1 into general subspace ΔW^g_k, whose signs align with those of the anchor and specific subspace ΔW^s_k, whose signs oppose those of the anchors. The sign-guided decomposition by anchor weights can be formulated as

$$\Delta W^g_k = M^g_k \odot \Delta W_k, \quad \Delta W^s_k = M^s_k \odot \Delta W_k, \tag{10}$$

where M^g_k and M^s_k denote binary masks computed by Eq. (6). The general subspace ΔW^g_k captures transferable updates that can be safely aggregated, while the specific subspace ΔW^s_k contains opposite-sign updates and therefore must be handled carefully to avoid knowledge collisions.

To mitigate this collision, we propose a selecting and scaling strategy. We first filter ΔW^s_k to retain only the most significant parameters (*e.g.*, those with large magnitudes), thereby removing small noise-like weights that may contribute to

interference. Then, we amplify the selected parameters using a scaling factor computed from the cosine distance $d_k = \delta(\Delta W_1^s, \Delta W_k^s)$ between selected specific parameters from ΔW_1^s and those from ΔW_k^s. Since the specific subspace has opposite signs to the anchor, the distance is empirically greater than 1 (*i.e.*, $d_k \geq 1$). Thus, the sign-guided LoRA merging is formulated as

$$\Delta W_k^s = \Delta W_k \odot \hat{M}_k^s, \text{ where } \hat{M}_k^s = M_k^s \wedge \mathbb{I}(|\Delta W_k| > \lambda |\Delta W_1|), \tag{11}$$

λ and $\hat{M}_k^s$ denote a temperature factor for anchor weights and the selected binary mask for a specific subspace. For inference, we unify all trained LoRA weights with the pre-trained weight W_θ. The final unified weight merging becomes

$$W' = W_\theta + \Delta W_1 + \sum_{k=2}^{t} (\Delta W_k^g + d_k \Delta W_k^s). \tag{12}$$

In summary, we identify salient parameters within the task-specific weights and scale them based on the degree of their specificity relative to the anchor weights. By suppressing the integration of insignificant weights while amplifying crucial task-specific weights, the proposed merging strategy effectively minimizes interference during the weight consolidation process. We summarize the overall algorithm of the SiGMA in the Appendix.

4 Experiments

We demonstrate that SiGMA effectively mitigates negative interference and subsequently improves the performance of MCIT across various benchmarks.

4.1 Experimental Settings

MCIT Benchmarks. We evaluate SiGMA on two behcnmarks: (1) UCIT (Unseen Continual Instruction Tuning) [8] includes six instruction-tuning datasets including ImageNet-R [13], ArixivQA [20], Vizwiz [11], IconQA [27], CLEVR-math [22], and Flicker30k [29]. They are chosen for LLaVA's weak zero-shot performance and filtered to avoid overlap with LLaVA's pre-training data, minimizing information leakage. (2) DCL (Domain Continual Learning) [39] can test the model trained sequentially across multiple domains, including remote sensing [25], medical [12], autonomous driving [32], science [4, 10, 18, 19], and finance [36, 39]. The severe domain shifts in the DCL benchmark often lead to strong cross-domain interference and significant forgetting.

Evaluation Metrics. Following standard continual learning evaluation protocols [8, 37], we report *Last*, *Average*, and *Forgets* to assess the performance. (1) *Last* is the accuracy on each previously seen task measured after training on the final task. (2) *Average* is reported in two forms; *Avg. All* is the mean of the last

UCIT	ImageNet	Arxiv	VizWiz	IconQA	CLEVR	Flicker30k	Avg. All	Forgets
Zero-shot	16.20	53.73	38.34	19.20	20.67	41.84	31.66	-
LoRA-FT [ICLR'22]	61.90	78.20	44.52	49.73	52.70	57.53	57.43	-17.02
O-LoRA [EMNLP'23]	72.73	77.77	43.87	45.83	55.33	57.27	58.80	-13.70
MoE LoRA [NeurIPS'24]	68.73	77.83	44.41	49.80	50.87	57.37	58.17	-15.24
CL-MoE [CVPR'25]	68.00	77.50	43.92	**52.37**	53.17	**57.80**	58.79	-15.29
HiDe LLaVA [ACL'25]	**84.70**	90.53	49.98	49.43	45.07	52.30	62.00	-5.58
DISCO [ICCV'25]	80.83	92.07	46.24	50.70	50.97	56.66	62.91	-6.27
SiGMA (Ours)	83.40	**94.07**	**54.83**	51.03	**57.83**	55.87	**66.17**	**-4.81**

DCL	RS	MED	AD	Sci	Fin	Avg. All	Forgets
Zero-shot	32.29	28.28	15.59	35.55	62.56	34.85	-
LoRA-FT [ICLR'22]	69.65	41.59	25.43	40.88	87.45	53.00	-14.97
O-LoRA [EMNLP'23]	76.04	44.71	36.60	46.35	90.19	58.78	-9.75
MoE LoRA [NeurIPS'24]	77.16	47.76	30.77	41.34	89.50	57.31	-12.19
CL-MoE [CVPR'25]	72.91	49.67	36.00	42.93	**90.22**	58.35	-11.56
HiDe LLaVA [ACL'25]	**78.84**	51.22	38.09	46.68	85.11	59.99	-3.86
DISCO [ICCV'25]	75.60	44.07	**51.68**	45.40	85.36	60.42	-5.44
SiGMA (Ours)	78.20	**59.75**	35.36	**50.73**	87.45	**62.30**	**-2.14**

Table 1: Comparison with recent methods on the UCIT (top) and DCL (bottom) benchmark in terms of Last, Avg. All, and Forgetting. The best and second-best results are highlighted in bold and underline, respectively.

accuracies across all tasks, and *Avg. Each* is the average accuracy evaluated immediately after the last task. (3) *Forgets* is the average drop in accuracy for each task, defined as the difference between its accuracy immediately after training and its final accuracy after the last task. In summary, *Avg. All*, *Avg. Each*, and *Forgets* respectively reflect overall knowledge retention, the ability to learn new tasks, and the extent of catastrophic forgetting in continual learning.

Implementation details. We use LLaVA-v1.5-7B [23] as the backbone model in all experiments. Following the LoRA fine-tuning setup from LLaVA, we insert LoRAs into the linear layers of the language model with a low-rank set to 16. We apply the proposed sign-guided adaptive tuning to all layers except the final output layers for next-token generation. For sign-guided LoRA merging, we set the temperature $\lambda = 0.8$. We set a batch size of 32 and 8 for all methods on UCIT and DCL benchmarks. We run all experiments on RTX A6000 GPUs.

4.2 Main Results

We compare SiGMA with state-of-the-art LoRA-based continual instruction tuning baselines such as LoRA-FT [14], O-LoRA [35], MoE LoRA [5], CL-MoE [15],

SiGMA		Metrics		
S-Tuning	S-Merge	Avg. All	Avg. Each	Forgets
		62.42	67.41	-5.98
	✓	63.35	68.57	-6.26
✓		64.86	68.56	**-4.44**
✓	✓	**66.17**	**70.18**	-4.81

(a) Ablation studies on SiGMA.

S-Tuning		Metrics		
LoRA Δ	SWD	Avg. All	Avg. Each	Forgets
$\sum_i B_i A_i$		63.23	67.85	-5.54
$\sum_i B_i \sum_i A_i$		62.42	67.41	-5.98
$\sum_i B_i A_i$	✓	64.15	**69.06**	-5.89
$\sum_\mathbf{i} \mathbf{B_i} \sum_\mathbf{i} \mathbf{A_i}$	✓	**64.86**	68.56	**-4.44**

(b) Analysis of S-Tuning.

Table 2: Ablation studies on S-Tuning and S-Merge of SiGMA (left), with analysis of S-Tuning (right). We examine the contribution of each component of SiGMA and analyze the effects of weight consolidation and sign-based weight decoupling (SWD).

S-Tuning	Metrics		
w/o top-k layers	Avg. All	Avg. Each	Forgets
k = 32	63.35	68.57	-6.26
k = 16	64.41	68.48	-4.88
k = 5	65.55	68.32	-5.38
k = 3	65.56	69.74	-5.01
k=1	**66.17**	**70.18**	**-4.81**
k = 0 (all layer)	64.27	68.49	-5.07

(a) Analysis of S-Tuning across the layer depth.

S-Merge	Metrics		
Temp λ (w/ d_k)	Avg. All	Avg. Each	Forgets
$\lambda = 0.$ (w/o λ)	54.32	70.69	-19.65
$\lambda = 0.6$	65.15	**70.56**	-6.49
$\lambda = 0.7$	65.88	70.40	-5.43
$\lambda = 0.8$ (w/o d_k)	62.69	67.11	-5.30
$\boldsymbol{\lambda = 0.8}$	**66.17**	70.18	-4.81
$\lambda = 0.9$	65.99	69.96	-4.76
$\lambda = 1.0$	65.89	69.81	**-4.70**

(b) Analysis of sign-guided LoRA merging.

Table 3: Analysis of S-Tuning and S-Merge. We analyze the effect of S-Tuning across layer depth (left) and the impact of the temperature factor λ and amplification term d_k in S-Merge (right) in terms of Average and Forgets.

Hide-LLaVA [8], and DISCO [9] on UCIT and DCL benchmarks. As shown in Table 1, SiGMA achieves the best Avg. All and Forgets on UCIT, indicating strong robustness to severe forgetting and negative interference. While HiDe-LLaVA and DISCO reduce forgetting via task-specific expansion and task-general fusion with auxiliary text encoders [30], SiGMA reports state-of-the-art performance without auxiliary models by mitigating interference at inference time. Notably, although all baselines score below 50.00% on VizWiz due to its difficulty and severe forgetting, SiGMA improves over the second-best method by +4.85%.

On the DCL benchmark, where each task requires domain-specific knowledge and the domain shifts are severe, prior methods suffer from severe forgetting, as shown in Table 1. In contrast, SiGMA achieves the best performance, outperforming the second-best method by +1.88% in Avg. All and +1.72% in Forgets. These results show that SiGMA preserves prior knowledge more effectively and remains robust under severe domain shift.

4.3 Further Analyses

To assess the effectiveness of SiGMA in more detail, we conduct additional experiments on the UCIT benchmarks. The proposed SiGMA has two main components: sign-guided adaptive tuning (coined as S-Tuning) and sign-guided LoRA merging (S-Merge). To gauge their individual contributions, we first perform ablations. As further analysis for S-Tuning, we test the effect of sign-based weight decoupling (SWD) and the layer position. For S-Merge, we ablate the temperature factor and evaluate the effects of amplification.

Ablation Study. We report the performance of ablations in Table 2a. As a baseline, we adopt an expansion–merge strategy that trains task-specific LoRA modules and merges all learned LoRAs at inference. Although this baseline can preserve prior weights to reduce severe forgetting, it remains susceptible to interference during consolidation. Applying S-Tuning or S-Merge alone still improves overall performance. However, S-Merge reports higher accuracy but slightly worse forgetting, since it does not explicitly enforce a discriminative separation between general and specific subspaces during training and thus may merge conflicting updates. S-Tuning consistently reduces forgetting by promoting more transferable updates that remain compatible when all task LoRAs are aggregated at inference. Combining S-Tuning and S-Merge achieves the best results, demonstrating their promising synergistic effect.

Analysis on Sign-guided Adaptive Tuning (S-Tuning). Table 2b analyzes the effects of sign-based weight decoupling (SWD) within S-Tuning and compares two consolidation strategies: merging the full LoRA updates Δ and merging the low-rank matrices A and B. SWD consistently improves overall performance under both strategies, supporting its role in conveying transferable knowledge from prior tasks. Notably, consolidating A and B is more effective than consolidating full LoRA weights Δ when SWD is used. While Δ-level consolidation simply combines task-specific knowledge, low-rank consolidation combines low-rank components across tasks, which better explores general latent subspaces. Combined with SWD, this provides a transferable subspace during training, reducing drift from previous tasks and improving adaptation to new tasks.

Furthermore, we investigate where to apply S-Tuning across layers. As shown in Table 3a, we increase the number of S-tuned layers by applying S-Tuning to all layers except the top-k layers of the LLM. Here, $k = 32$ means absence of S-Tuning, and $k = 0$ means S-Tuning applied to all layers. Overall performance increases as S-Tuning is applied to more layers, except when all layers are tuned. This performance drop is expected because the top-1 layer directly connects to the next-token prediction, injecting overly task-invariant knowledge can degrade discriminative decisions. Accordingly, excluding the top-1 layer shows strong Avg. All and Avg. Each scores and the lowest Forgets.

Analysis on Sign-guided LoRA Merging (S-Merge). We first examine how the varied temperature factor λ affects the selection of salient parameters.

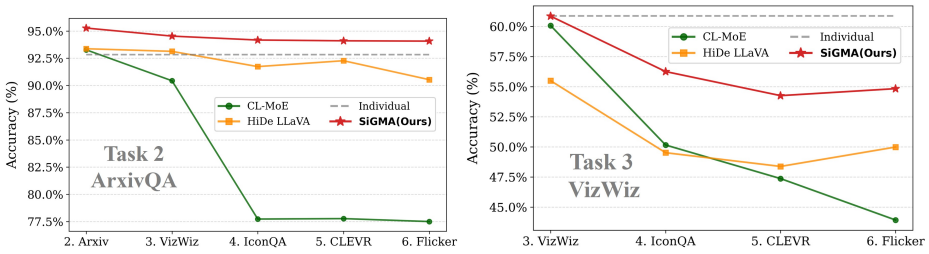

Fig. 4: Performance variation under the UCIT benchmark. We report changes in single-task performance (*e.g.*, ArxivQA and VizWiz) as the model sequentially learns new tasks. "Individual" denotes the performance achieved when the model is trained only on the corresponding task.

As shown in Table 3b, increasing the temperature λ improves overall performance up to 0.8. Although the best Avg. Each and Forgets report at slightly different temperatures, we set $\lambda = 0.8$ because it best reflects overall knowledge retention. Using the optimal temperature, we then test the effects of weight amplification by cosine distance d_k. Removing d_k causes a clear degradation: Avg. All and Avg. Each drop by 3.48% and 3.07%, respectively, and Forgets increases by 0.49%. These results demonstrate that amplifying selected parameters helps preserve their knowledge when consolidating weights at inference.

As λ increases, we observe a clear trade-off; Avg. Each slightly declines, while Forgets improves significantly. Specifically, the degradation in Avg. Each is marginal (dropping from 70.69% to 69.81%), whereas the reduction in forgetting is substantial (improving from -19.65% to -4.70). These results demonstrate that enforcing stricter parameter selection, which restricts the number of specific parameters merged during inference, sacrifices only a minimal amount of plasticity (performance on the current task) in exchange for a substantial gain in stability. Intuitively, S-Merge enables the weights to filter out less critical parameters, preventing negative interference while preserving prior knowledge.

Analysis on Negative Interference. To validate mitigated negative interference, we track the performance of specific tasks (*e.g.*, ArxivQA and VizWiz) as the model sequentially learns subsequent tasks. As shown in Figure 4, prior methods still suffer from severe degradation due to negative interference. Specifically, CL-MoE shows a drastic decline as new tasks are introduced, and HiDe LLaVA, despite preserving prior weights intact, reports sub-optimal performance. In contrast, SiGMA achieves outstanding performance in preserving prior knowledge. Notably, in ArxivQA, SiGMA consistently outperforms individual training throughout the entire learning process. This shows that SiGMA effectively mitigates negative interference, improving positive transferability.

Qualitative Results. Figure 5 shows qualitative examples on the UCIT benchmark, comparing SiGMA with state-of-the-art methods: CL-MoE and HiDE-

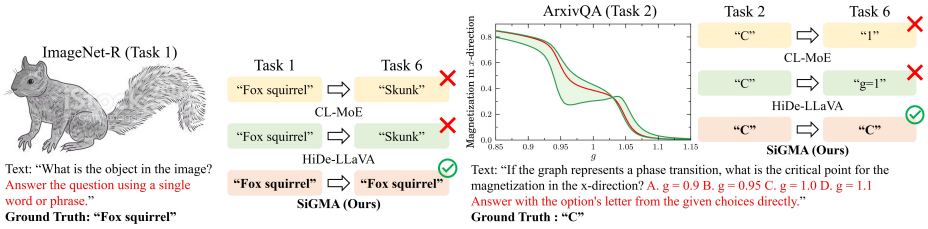

Fig. 5: Qualitative examples from ImageNet-R (left) and ArxivQA (right) on the UCIT benchmark. Given the same image and instruction, we compare responses from CL-MoE, HiDe-LLaVA, and SiGMA, shown both after task-specific training and after sequential training on all tasks. Red text denotes the task-specific instruction format that the model must follow to generate a response.

LLaVA. All methods follow instructions and produce correct answers after training on the corresponding task. However, after sequentially training all tasks, CL-MoE and HiDE-LLaVA suffer from generating incorrect responses on ImageNet-R (*e.g.*, misclassifying a "Fox squirrel" as a "Skunk") and failing instruction following (*e.g.*, generating the full content of an answer option instead of simply producing the corresponding option letter). In contrast, SiGMA maintains strong instruction-following ability and preserves correct task knowledge. By effectively mitigating negative interference, it ensures that the model reliably adapts to new, task-specific instructions without suffering from catastrophic forgetting.

5 Conclusion

We address negative interference, a critical yet often overlooked challenge in MCIT, where the consolidation of new weights disrupts previously acquired knowledge. Unlike prior methods that primarily focus on catastrophic forgetting, SiGMA explicitly aims to mitigate negative interference. Based on sign-based weight decoupling, our sign-guided adaptive tuning enables stable and transferable learning with minimal parameter drift, while sign-guided LoRA merging mitigates negative interference by selectively amplifying salient parameters.

Despite its strong performance, SiGMA still has room for improvement. The general subspace may become increasingly sparse as the task sequence grows, reducing its effectiveness in long-term continual instruction tuning, for which future work will explore strategies to robustly maintain the general subspace. Furthermore, we plan to extend sign-based weight decoupling beyond LoRA to other Parameter-Efficient Fine-Tuning (PEFT) techniques, such as adapters and learnable prompts. Given their widespread adoption for efficient adaptation, we believe that extending sign-based interference across diverse PEFT architectures will shed light on broader insights into the continual learning paradigm.

Acknowledgements

This work was supported by Samsung Electronics Co., Ltd (IO250418-12669-01) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT): (1) No. RS-2026-25524173, Ultro-Long-Term Hierarchical Memory and Reasoning Architecture for Next-Generation Omnimodal Agents, (2) No. RS-2019-II191082, SW StarLab, (3) No. RS-2022-II220156, Fundamental research on continual meta-learning for quality enhancement of casual videos and their 3D metaverse transformation, and (4) No. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University). Gunhee Kim is the corresponding author.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Aljundi, R., Belilovsky, E., Tuytelaars, T., Charlin, L., Caccia, M., Lin, M., Page-Caccia, L.: Online continual learning with maximal interfered retrieval. Advances in neural information processing systems **32** (2019)
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
4. Chang, S., Palzer, D., Li, J., Fosler-Lussier, E., Xiao, N.: Mapqa: A dataset for question answering on choropleth maps. arXiv preprint arXiv:2211.08545 (2022)
5. Chen, C., Zhu, J., Luo, X., Shen, H.T., Song, J., Gao, L.: Coin: A benchmark of continual instruction tuning for multimodel large language models. Advances in Neural Information Processing Systems **37**, 57817–57840 (2024)
6. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
7. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems **36**, 49250–49267 (2023)
8. Guo, H., Zeng, F., Xiang, Z., Zhu, F., Wang, D.H., Zhang, X.Y., Liu, C.L.: HiDe-LLaVA: Hierarchical decoupling for continual instruction tuning of multimodal large language model. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 13572–13586. Association for Computational Linguistics, Vienna, Austria (Jul 2025). `https://doi.org/10.18653/v1/2025.acl-long.666`, `https://aclanthology.org/2025.acl-long.666/`
9. Guo, H., Zeng, F., Zhu, F., Liu, W., Wang, D.H., Xu, J., Zhang, X.Y., Liu, C.L.: Federated continual instruction tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1325–1335 (2025)
10. Guo, Z., Zhang, R., Chen, H., Gao, J., Jiang, D., Wang, J., Heng, P.A.: Sciverse: Unveiling the knowledge comprehension and visual reasoning of lmms on multimodal scientific problems. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 19683–19704 (2025)

11. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: 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)
12. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
13. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8340–8349 (2021)
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9
15. Huai, T., Zhou, J., Wu, X., Chen, Q., Bai, Q., Zhou, Z., He, L.: Cl-moe: Enhancing multimodal large language model with dual momentum mixture-of-experts for continual visual question answering. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19608–19617 (2025)
16. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural computation **3**(1), 79–87 (1991)
17. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b. ArXiv **abs/2310.06825** (2023), https://api.semanticscholar.org/CorpusID:263830494
18. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: European conference on computer vision. pp. 235–251. Springer (2016)
19. Kembhavi, A., Seo, M., Schwenk, D., Choi, J., Farhadi, A., Hajishirzi, H.: Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In: Proceedings of the IEEE Conference on Computer Vision and Pattern recognition. pp. 4999–5007 (2017)
20. Li, L., Wang, Y., Xu, R., Wang, P., Feng, X., Kong, L., Liu, Q.: Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 14369–14387 (2024)
21. Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y., Bai, X.: Monkey: Image resolution and text label are important things for large multimodal models. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26763–26773 (2024)
22. Lindström, A., Abraham, S.: Clevr-math: A dataset for compositional language, visual and mathematical reasoning. In: Proceedings of the 16th International Workshop on Neural-Symbolic Learning and Reasoning. vol. 3212, pp. 155–170 (2022)
23. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
24. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
25. Lobry, S., Marcos, D., Murray, J., Tuia, D.: Rsvqa: Visual question answering for remote sensing data. IEEE Transactions on Geoscience and Remote Sensing **58**(12), 8555–8566 (2020)

26. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In: The Twelfth International Conference on Learning Representations (2024)
27. Lu, P., Qiu, L., Chen, J., Xia, T., Zhao, Y., Zhang, W., Yu, Z., Liang, X., Zhu, S.C.: Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. arXiv preprint arXiv:2110.13214 (2021)
28. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. In: Psychology of learning and motivation, vol. 24, pp. 109–165. Elsevier (1989)
29. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
31. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In: International Conference on Learning Representations (2017)
32. Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beißwenger, J., Luo, P., Geiger, A., Li, H.: Drivelm: Driving with graph visual question answering. In: European conference on computer vision. pp. 256–274. Springer (2024)
33. Tang, B., Boggust, A., Satyanarayan, A.: Vistext: A benchmark for semantically rich chart captioning. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7268–7298 (2023)
34. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
35. Wang, X., Chen, T., Ge, Q., Xia, H., Bao, R., Zheng, R., Zhang, Q., Gui, T., Huang, X.J.: Orthogonal subspace learning for language model continual learning. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 10658–10671 (2023)
36. Wang, Z., Li, Y., Wu, J., Soon, J., Zhang, X.: Finvis-gpt: A multimodal large language model for financial chart analysis. arXiv preprint arXiv:2308.01430 (2023)
37. Wang, Z., Zhang, Z., Lee, C.Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., Pfister, T.: Learning to prompt for continual learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 139–149 (2022)
38. Zeng, F., Zhu, F., Guo, H., Zhang, X.Y., Liu, C.L.: Modalprompt: Towards efficient multimodal continual instruction tuning with dual-modality guided prompt. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 12137–12152 (2025)
39. Zhao, H., Zhu, F., Guo, H., Wang, M., Wang, R., Meng, G., Zhang, Z.: Mllm-cl: Continual learning for multimodal large language models. arXiv preprint arXiv:2506.05453 (2025)
40. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)