Progressive Multimodal Alignment for Continual Instruction Tuning

Duzhen Zhang*
Mohamed bin Zayed University of Artificial Intelligence
Abu Dhabi, United Arab Emirates
Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences
Shanghai, China
duzhen.zhang@mbzuai.ac.ae

Yahan Yu*
Kyoto University
Kyoto, Japan
yahan@nlp.ist.i.kyoto-u.ac.jp

Qiaoyi Su
Migu Culture Technology Co.,Ltd.
Beijing, China
suqiaoyi@migu.chinamobile.com

Jiahua Dong
Mohamed bin Zayed University of Artificial Intelligence
Abu Dhabi, United Arab Emirates
dongjiahua1995@gmail.com

Tielin Zhang†
Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences
Shanghai, China
State Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology
Shanghai, China
zhangtielin@ion.ac.cn

Abstract

Multimodal Large Language Models (MLLMs) rely on a projector to align visual representations with the language embedding space, making it central to cross-modal understanding. In Multimodal Continual Instruction Tuning (MCIT), however, shifting visual distributions and evolving instruction semantics cause this shared projector to drift, leading to projector-level forgetting, an issue largely overlooked by methods that focus primarily on the LLM backbone. We introduce Progressive Multimodal Alignment (PMA), a framework that enables the projector to adapt continually while preserving previously learned alignment. PMA detects multimodal distribution shifts via a lightweight representation descriptor and progressively expands projector experts only when needed. An expandable router integrates expert outputs based on multimodal features, while the original pretrained projector is retained as a stable alignment anchor. This progressive mechanism balances stability and plasticity with sub-linear parameter growth and serves as a method-agnostic add-on to existing MCIT approaches. Extensive experiments on two recent MCIT benchmarks demonstrate that mitigating projector-level forgetting yields consistent gains over prior state-of-the-art methods when combined with PMA. Moreover, PMA scales across diverse MLLM backbones, demonstrating robust and broadly applicable MCIT performance.[1]

CCS Concepts

• **Computing methodologies** → **Lifelong machine learning**.

Keywords

Multimodal Large Language Models, Multimodal Continual Instruction Tuning, Multimodal Alignment

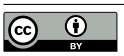
ACM Reference Format:
Duzhen Zhang, Yahan Yu, Qiaoyi Su, Jiahua Dong, and Tielin Zhang. 2026. Progressive Multimodal Alignment for Continual Instruction Tuning. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3836259

1 Introduction

Multimodal Large Language Models (MLLMs) have advanced visual-language understanding and instruction following by coupling strong visual encoders with LLMs capable of open-ended reasoning [31, 46, 47]. A central component in this architecture is the projector, which maps visual features into the language embedding space and serves as the semantic interface for cross-modal alignment.

As MLLMs are increasingly deployed in dynamic environments, they must adapt to evolving tasks and instruction styles rather than operate as static, once-trained models. However, retraining them

*Both authors contributed equally to this research.
†Corresponding author.



MM '26, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3836259

[1]The code is available at https://github.com/BladeDancer957/PMA.

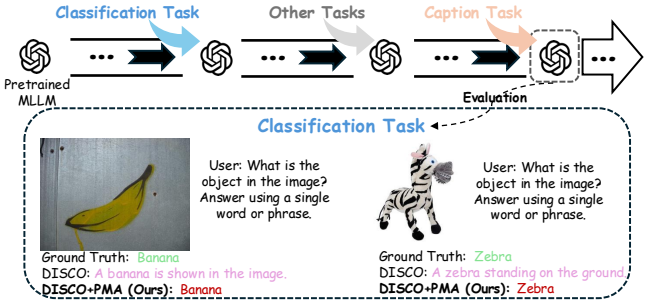


Figure 1: Illustration of projector-level forgetting in MCIT. After finetuning on a captioning task, DISCO [17] produces caption-style responses even for classification instructions, indicating degraded visual translation for earlier tasks. In contrast, DISCO+PMA (Ours) preserves task-specific cross-modal alignment and generates task-consistent classification outputs for the same inputs.

for every new task is prohibitively costly, motivating growing interest in Multimodal Continual Instruction Tuning (MCIT) [3, 5, 21], where models aim to acquire new capabilities while preserving both previously learned skills and established cross-modal alignment. Achieving this balance is challenging due to Catastrophic Forgetting (CF) [10, 11, 15, 28, 40], as shifts in visual distributions and instruction semantics can cause models to lose prior knowledge and misalign earlier visual-language mappings.

To mitigate CF, existing MCIT methods primarily incorporate Parameter-Efficient FineTuning (PEFT) techniques, such as prompt tuning [29, 49] and LoRA [24, 48], into the LLM backbone to preserve previously acquired capabilities. For instance, MoELoRA [5] employs multi-expert LoRA architectures [12, 37, 50] to capture task-specific knowledge; HiDE [16] leverages layer-wise similarity variations to decompose LoRA into task-specific expansion and task-general fusion components, striking a balance between adaptation performance and memory efficiency; and DISCO [17] assigns task-specific LoRA subspaces with subspace-selective activation to reduce interference, achieving State-Of-The-Art (SOTA) MCIT performance. However, these approaches largely overlook the preservation of cross-modal alignment learned in earlier tasks, implicitly treating the projector as a shared, jointly finetuned module across all tasks. As a result, updating this shared projector induces alignment drift, whereby newly learned tasks overwrite previously established vision-language mappings, a phenomenon we term *projector-level forgetting*.

Taking DISCO [17] as an illustrative example, after finetuning on an image captioning task, the model may exhibit a bias toward the most recent instruction style, causing incorrect translation of visual features for earlier image classification tasks. As illustrated in Figure 1, we observe cases where the model produces caption-style responses even when instructed to output a concise classification label. This behavior reflects an instruction-driven collapse in visual translation, where the projector fails to condition its mappings on earlier task instructions, leading to degraded cross-modal alignment and projector-level forgetting. In contrast, when equipped with our method, the same inputs are routed to appropriate projector experts, and the model generates task-consistent classification outputs, avoiding this failure mode.

A naive solution is to allocate a separate projector for each task. However, this strategy is parameter-inefficient, scales linearly with the number of tasks, hinders knowledge sharing among related tasks, and requires task IDs at inference, which is incompatible with realistic MCIT settings where task IDs are unavailable. This leads to a fundamental question: *How can we enable the projector to preserve its original alignment while adapting to new multimodal tasks in a parameter-efficient, transferable, and task-ID-free manner?*

In this paper, we address this challenge with Progressive Multimodal Alignment (PMA), a framework that allows the projector to adapt continually while retaining previously learned alignment. PMA detects multimodal distribution shifts via a lightweight Representation Descriptor (RD) and expands new projector experts only when necessary, promoting knowledge sharing and ensuring sub-linear parameter growth. An expandable router dynamically integrates expert outputs based on the same multimodal features as RD, enabling automatic routing without task-ID supervision, while the frozen pretrained projector provides a stable alignment anchor. Importantly, PMA is method-agnostic and integrates seamlessly with existing MCIT approaches, complementing their focus on the LLM backbone by directly addressing the long-overlooked issue of projector-level forgetting.

Our contributions can be summarized as follows:

- We identify and formalize *projector-level forgetting* as a key yet largely overlooked bottleneck in MCIT, demonstrating that drift in the projector undermines instruction retention.
- We propose PMA, a method-agnostic framework that enables the projector to adapt continually by detecting multimodal distribution shifts, selectively expanding lightweight experts, and dynamically routing them without task IDs, all while preserving previously learned alignment with sub-linear parameter growth.
- We conduct extensive experiments on two MCIT benchmarks (UCIT [16] and MLLM-DCL [53]), showing that PMA consistently improves previous SOTA methods and scales effectively across diverse MLLM backbones (LLaVA-1.5 [34], InternVL [9]).

2 Related Work

2.1 MLLMs

Recent progress in MLLMs has significantly advanced visual-language understanding [6] and instruction following [51]. Early models such as BLIP-2 [31] employ a frozen LLM paired with a frozen visual encoder and a learnable projector (*e.g.*, Q-Former) to achieve efficient modality alignment. Subsequent systems, including LLaVA [36], MiniGPT-4 [56], and QwenVL [1], simplify the alignment mechanism using linear projectors and show that instruction tuning plays a crucial role in aligning visual features with human intent. Recent variants, such as LLaVA-1.5 [34], ShareGPT4V [8], and InternVL [9], further refine these alignment strategies and show strong performance across a diverse set of multimodal benchmarks. Meanwhile, the MLLM ecosystem has expanded beyond static images to modalities such as video and audio [2, 13, 30, 35, 41], signaling a broader

shift toward more general-purpose multimodal reasoning. However, as model scale and application complexity continue to grow, adapting MLLMs to evolving tasks and instruction styles without retraining from scratch becomes both necessary and challenging. This demands new paradigms for MCIT, enabling MLLMs to maintain alignment with human intent in dynamic, real-world environments.

2.2 MCIT

Building on the need for adaptable MLLMs, recent work has begun to explore MCIT, which aims to maintain alignment as instruction styles and task distributions evolve. To support systematic evaluation, several benchmarks have been proposed [3, 21]. CoIN [5] and UCIT [16] introduce dataset-incremental settings; however, CoIN suffers from pretraining overlap that leads to information leakage, while UCIT addresses this by selecting datasets minimally correlated with LLaVA's pretraining data. More recent efforts like MLLM-DCL [53] further broaden the benchmark landscape with domain-specific knowledge.

To mitigate CF, recent research work adapts various PEFT strategies to the MCIT setting [14, 52, 54, 55]. MCITlib [18] provides a unified framework that consolidates representative approaches such as MoELoRA [5] maintains multiple LoRA experts to capture task-specific information; SEFE [7] addresses both superficial and essential forgetting by harmonizing task styles via answer style diversification and stabilizing critical parameters with RegLoRA; and DISCO [17] allocates task-specific LoRA subspaces during training and employs subspace-selective activation during inference to reduce interference. More method introductions are provided in Section 4.2.

While existing MCIT methods primarily address CF in the LLM backbone, projector-level forgetting remains largely overlooked. To address this gap, we propose PMA, which expands representational capacity only when needed while preserving previously learned alignment. By directly targeting projector-level forgetting, an underexplored but critical bottleneck in MCIT, PMA integrates seamlessly with nearly all existing PEFT-based MCIT methods.

3 Method

We propose PMA, a method-agnostic framework that complements existing MCIT approaches by explicitly addressing projector-level forgetting, an issue largely overlooked by methods that focus on the LLM backbone. As illustrated in Figure 2, PMA employs lightweight RDs to detect multimodal distribution shifts and expands the projector with a new expert only when such shifts are detected. An expandable router integrates expert outputs based on the same multimodal features used by the RD, enabling task-agnostic inference and facilitating knowledge sharing across related tasks, while the frozen pretrained projector is retained as a stable alignment anchor. Overall, PMA provides a progressive and parameter-efficient mechanism for maintaining cross-modal alignment, and can be seamlessly integrated with prior PEFT-based MCIT methods.

3.1 Task Formulation

MCIT [18] aims to update an MLLM with new instruction-driven tasks without incurring the cost of full retraining. We consider a setting where an MLLM is finetuned over a sequence of tasks $t = 1, \ldots, T$, each associated with a training set $\mathcal{D}^t_{\text{train}}$ and a test set $\mathcal{D}^t_{\text{test}}$. Each instance $x^{t,i}$ in these datasets consists of an image $x^{t,i}_{\text{img}}$, a text instruction $x^{t,i}_{\text{txt}}$, and an answer $x^{t,i}_{\text{ans}}$. The goal is to incrementally adapt a single model $\mathcal{M}$ while preserving strong performance on all previously learned tasks. MCIT is typically evaluated in a **rehearsal-free** setting, where data from earlier tasks cannot be revisited during later training, and task identities remain **unknown** at inference time.

3.2 Initialization for the First Task

For the first task t=1, PMA initializes a projector $\mathcal{P}^1$ together with two associated components: a lightweight Representation Descriptor $\mathcal{RD}^1$, implemented as an Multi-Layer Perceptron (MLP)-based autoencoder, and an initial router $\mathcal{R}^1$. These components form the foundation for all subsequent progressive expansions.

For each training instance $x^{1,i} \in \mathcal{D}^1_{\text{train}}$, we first feed its image $x^{1,i}_{\text{img}}$ into the frozen visual encoder to obtain visual token embeddings, which are averaged to produce a global visual feature $\boldsymbol{x}^{1,i}_{\text{img}} \in \mathbb{R}^{d_1}$. The initial projector $\mathcal{P}^1$ maps this feature into the language embedding space:

$$\tilde{\boldsymbol{x}}^{1,i}_{\text{img}} = w^1 \mathcal{P}^1(\boldsymbol{x}^{1,i}_{\text{img}}), \tag{1}$$

where w^1 represents the routing weight associated with $\mathcal{P}^1$. The router $\mathcal{R}^1$ computes this weight from a unified multimodal representation:

$$\begin{aligned} w^1 &= \text{Softmax}(\mathcal{R}^1(\boldsymbol{x}^{1,i}_{\text{fuse}})) \\ \mathcal{R}^1(\boldsymbol{x}^{1,i}_{\text{fuse}}) &= \boldsymbol{W}^{1^T} \cdot \boldsymbol{x}^{1,i}_{\text{fuse}} \end{aligned}, \tag{2}$$

where $\boldsymbol{W}^1$ is the learnable weight matrix of $\mathcal{R}^1$ and $\boldsymbol{x}^{1,i}_{\text{fuse}} \in \mathbb{R}^{d_1+d_2} = [\boldsymbol{x}^{1,i}_{\text{img}}; \boldsymbol{x}^{1,i}_{\text{txt}}]$ concatenates the global visual feature with the averaged instruction-token embedding $\boldsymbol{x}^{1,i}_{\text{txt}} \in \mathbb{R}^{d_2}$. This fused representation enables routing decisions to depend on both visual content and instruction semantics, which is essential for MCIT. At t=1, PMA instantiates a default projector expert $\mathcal{P}^1$ together with the initial router $\mathcal{R}^1$. Since only a single expert is available, the matrix $\boldsymbol{W}^1$ has one column, and the Softmax degenerates to a constant selection, yielding w^1=1.0 for all samples.

$\mathcal{RD}^1$ operates on the same fused representation as $\mathcal{R}^1$, allowing it to model both the visual distribution and instruction semantics. Implemented as a small autoencoder, $\mathcal{RD}^1$ takes $\boldsymbol{x}^{1,i}_{\text{fuse}}$ as input and reconstructs it as $\hat{\boldsymbol{x}}^{1,i}_{\text{fuse}}$. The reconstruction error

$$r = ||\boldsymbol{x}^{1,i}_{\text{fuse}} - \hat{\boldsymbol{x}}^{1,i}_{\text{fuse}}||_2^2 \tag{3}$$

serves as a measure of how well the current projector configuration accounts for the new task.

After training $\mathcal{RD}^1$ on the first task, we compute the mean μ^1 and standard deviation σ^1 of reconstruction errors across all training samples. These statistics define a reference distribution that characterizes the multimodal patterns of task t=1. For subsequent tasks, reconstruction errors produced by $\mathcal{RD}^1$ are compared against this baseline to determine whether incoming representations deviate substantially from those seen in the first task. This lightweight, data-driven mechanism allows PMA to detect task novelty and ensures that projector expansion is triggered only when necessary.

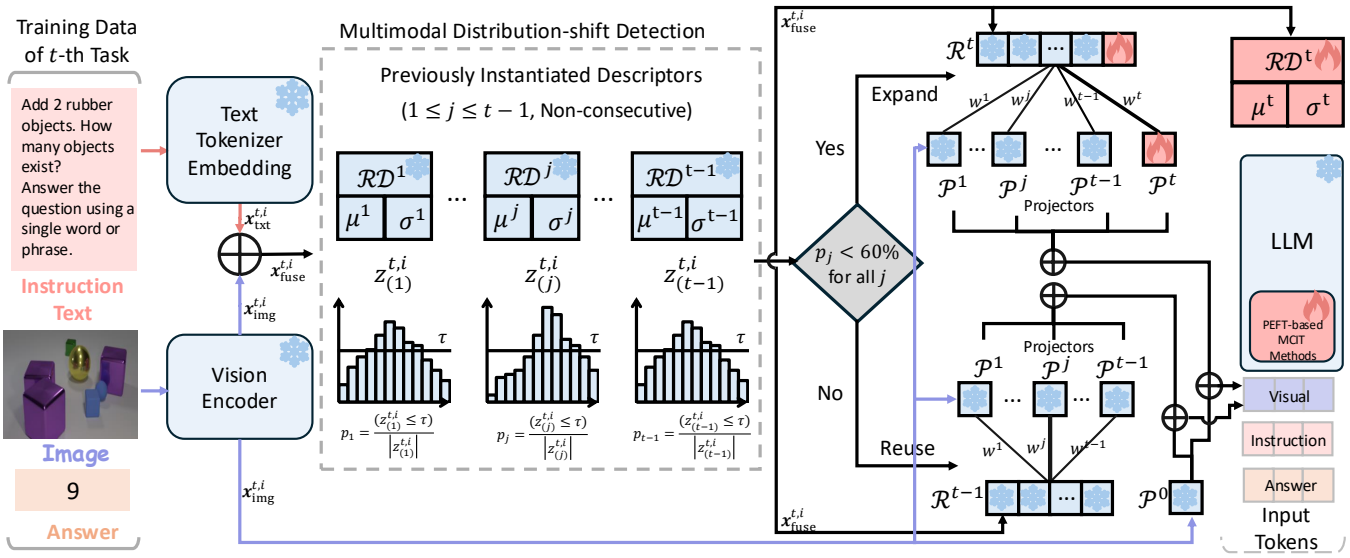


Figure 2: Overview of PMA. Lightweight RDs detect multimodal distribution shifts and determine whether to trigger projector expansion or reuse existing projector experts. An expandable router ($\mathcal{R}^t$) mixes projector experts without task IDs, while the frozen pre-trained projector ($\mathcal{P}^0$) serves as a stable alignment anchor. PMA is method-agnostic and integrates seamlessly with existing PEFT-based MCIT methods that primarily focus on the LLM side, enabling progressive projector-side adaptation with sub-linear parameter growth.

3.3 Expansion for Subsequent Tasks

For each subsequent task $t \geq 2$, PMA determines whether the existing projector experts can adequately model the new multimodal representations. Given the fused representation $\boldsymbol{x}^{t,i}_{\text{fuse}}$ of an instance i from task t, PMA evaluates it against **all previously instantiated descriptors** $\{\mathcal{RD}^j \mid 1 \leq j \leq t-1,\ \mathcal{RD}^j \text{ exists}\}$. Each $\mathcal{RD}^j$ produces reconstruction errors $r^{t,i}_{(j)}$, which are standardized using the statistics (μ^j, σ^j) collected from task j:

$$z^{t,i}_{(j)} = \frac{r^{t,i}_{(j)} - \mu^j}{\sigma^j}. \tag{4}$$

We rely on z-scores rather than raw reconstruction errors, as they normalize scale differences across descriptors and tasks, making deviations from prior distributions comparable and robust. Each sample thus obtains a z-score for each prior descriptor. For each $\mathcal{RD}^j$, PMA computes the proportion of samples p_j whose z-scores satisfy $z^{t,i}_{(j)} \leq \tau$.

If **all descriptors** yield proportions $p_j < 60\%$, PMA concludes that the new task introduces a multimodal distribution not captured by existing experts and allocates a new projector expert $\mathcal{P}^t$, a corresponding descriptor $\mathcal{RD}^t$, and a new weight column in $\mathcal{R}^t$. **Only** these newly added components are updated for task t, which biases the router toward assigning higher weights to the new projector on this task; we then compute the mean μ^t and standard deviation σ^t of the reconstruction errors. All previously learned projectors, descriptors, and old router columns remain frozen.

Otherwise, PMA determines that no expansion is required and proceeds without introducing new components. In this case, PMA reuses all existing projector experts together with their corresponding router weight columns. Crucially, during reuse, **all** projectors, descriptors, and router columns **remain frozen** and are not updated using data from task t, thereby preventing interference with previously learned tasks. To identify the most relevant prior knowledge, PMA selects $j^\star = \arg\max_j p_j$, corresponding to the most compatible prior task. Since the projector $\mathcal{P}^{j^\star}$ was trained to receive dominant routing weights on its originating task and the router parameters remain frozen thereafter, and since both the descriptor and the router operate on the same fused multimodal representation, the router is encouraged to assign a larger routing weight to $\mathcal{P}^{j^\star}$ when processing a new task with a similar multimodal distribution, while assigning smaller but non-zero weights to other projector experts.

This expansion mechanism promotes knowledge sharing across related tasks and ensures sub-linear parameter growth, as new experts are introduced only when PMA detects a genuinely novel multimodal distribution not explained by any previous RD. PMA thus provides a principled, data-driven mechanism for deciding when to reuse or expand projector capacity, enabling continual multimodal adaptation while maintaining stable visual-language alignment across tasks.

Throughout training, on both the first and subsequent tasks, the frozen pretrained projector $\mathcal{P}^0$ is kept as a stable alignment anchor, enabling PMA to preserve the core visual-language mapping established during pretraining while progressively adapting to new tasks. The final projected representation is given by

$$\tilde{\boldsymbol{x}}^{t,i}_{\text{img}} = \frac{1}{1+t'}\left(\mathcal{P}^0(\boldsymbol{x}^{t,i}_{\text{img}}) + \sum_{j=1}^{t'} w^j \mathcal{P}^j(\boldsymbol{x}^{t,i}_{\text{img}})\right), \tag{5}$$

where t' is the number of instantiated projector experts ($1 \leq t' < t$, ensuring sub-linear growth), and w^j is the routing weight assigned

to expert j by the router $\mathcal{R}^t$, computed from the fused representation $\boldsymbol{x}_{\text{fuse}}^{t,i}$ via Equation (2). The resulting projected representation, aligned with the LLM's language embedding space, is fused with the instruction embeddings and passed to the LLM to generate the final prediction.

3.4 Training Objective and Inference

PMA optimizes two largely independent components: (1) the cross-modal alignment pathway, consisting of the projector experts and the router, and (2) the task-specific descriptor associated with each expert. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \mathcal{L}_{\text{RD}}, \tag{6}$$

where $\mathcal{L}_{\text{LM}}$ is the autoregressive language modeling loss used for instruction tuning, and $\mathcal{L}_{\text{RD}}$ (Equation (3)) is the reconstruction loss. These losses are fully decoupled: $\mathcal{L}_{\text{LM}}$ updates only the active projector expert and router column selected by PMA, along with the LLM-side PEFT parameters (depending on the combined MCIT method), whereas $\mathcal{L}_{\text{RD}}$ trains the descriptor alone and receives no gradient from the language modeling objective. This separation allows PMA to remain method-agnostic and integrate seamlessly with prior MCIT methods that mainly focus on mitigating CF in the LLM backbone.

4 Experimental Settings

4.1 Datasets and Benchmarks

As the MLLMs have already seen large-scale image-text pairs during pretraining, we adopt two benchmarks from MCITlib [18] designed to mitigate information leakage in MCIT training: (1) **UCIT** benchmark [16] consists of ImageNet-R (ImgNet-R) [23], ArxivQA [32], VizWiz-Caption (VizWiz) [20], IconQA [39], CLEVR-Math (CLEVR) [33], and Flickr30k [42], including image captioning, Visual Question Answering (VQA), and multiple-choice reasoning tasks. The MLLMs exhibit weak zero-shot performance, suggesting a low risk of information leakage. All datasets are trained in the above order as in the main experiments. (2) **MLLM-DCL** benchmark [53] extends to downstream tasks from five domains—Remote Sensing (RS), Medicine (Med), Autonomous Driving (AD), Science (Sci), and Finance (Fin)—trained in the order RS → Med → AD → Sci → Fin, incorporating RSVQA [38], PathVQA [22], DriveLM [43], AI2D [26], Sciverse [19], MapQA [4], TQA [27], and FinVis [45] datasets.

4.2 Comparison Baselines

We compare our method against a set of representative baselines:

- **LoRA-FT** [24]: Sequentially updates knowledge through shared low-rank matrices while keeping the pretrained MLLM parameters frozen.
- **OLoRA** [44]: Mitigates CF by assigning each task to an orthogonal subspace, reducing interference across tasks.
- **MoELoRA** [5]: Employs multiple independent LoRAs to capture task-specific knowledge from sequential training.
- **CL-MoE** [25]: Adopts a dual-momentum MoE framework that dynamically selects and updates global and local experts through task-level and instance-level routers, enabling MCIT without CF.
- **SEFE** [7]: Tackles two types of forgetting—superficial and essential—by harmonizing task styles via answer style diversification and applying RegLoRA regularization to stabilize key parameters.
- **HiDE** [16]: Designs a task-specific LoRA expansion and task-general LoRA fusion strategy leveraging layer-wise similarity analysis to balance performance and efficiency while maintaining low memory usage.
- **DISCO** [17]: Introduces a dynamic knowledge organization mechanism that allocates task-specific LoRA subspaces, with sharing among related tasks, during training, and employs subspace-selective activation at inference time to mitigate cross-task interference.

In addition, we consider two reference settings for comparison: **Zero-shot**: Evaluates each task directly using the pretrained MLLM without additional finetuning. **Individual**: Finetunes the MLLM with LoRA independently on each downstream task, producing a distinct model per task without shared parameters.

4.3 Evaluation Metrics

Following MCITlib [18], we assess MCIT performance with four complementary evaluation metrics. **Mean Finetune Accuracy (MFT)** reports the average accuracy obtained on each task immediately after training, reflecting the model's learning ability without the influence of forgetting. **Mean Final Accuracy (MFN)** averages the accuracies of all tasks after the final training stage, indicating how well knowledge is retained overall. **Mean Average Accuracy (MAA)** takes the mean of the averaged accuracies across all intermediate training steps, providing a comprehensive view of performance evolution. **Backward Transfer (BWT)** measures the accuracy difference between the final and post-training states of each task, quantifying the degree of forgetting.

4.4 Implementation Details

As PMA is model-agnostic, we incorporate it into two representative MCIT baselines—HiDE and DISCO—to enhance their effectiveness. All baselines are built upon widely used MLLMs, including **LLaVA-1.5-7B** [34] and **InternVL-Chat-7B** [9], and are trained with LoRA. The vision encoder and LLM are frozen, while only the projector and LoRA modules are updated. The fused representation has dimension d_1+d_2, where the CLIP visual embedding dimension is d_1=768, and the LLM embedding dimension d_2 is 4096 for the 7B backbone. Each RD is implemented as a shallow MLP autoencoder with a single bottleneck layer of size $(d_1+d_2)/4$, trained for 1 epoch with a learning rate of $1e$-4. We set the z-score threshold to τ=1.4 for distribution-shift detection. All experiments are conducted on 4 NVIDIA A100 GPUs, each with 80 GB of memory.

5 Experimental Results

5.1 Main Results

We evaluate PMA on two representative MCIT benchmarks, UCIT and MLLM-DCL, covering diverse task types, instruction styles, and visual distributions. Tables 1, 2, 3, and 4 report the results on LLaVA-1.5-7B and InternVL-Chat-7B backbones across these benchmarks,

Table 1: Main results of LLaVA-1.5-7B on UCIT. The middle columns for each task report performance after finetuning on the final task. The bold denotes the highest result. ∗ denotes results from our re-implementation; all other numbers are taken from MCITlib [18]. HiDE and DISCO with PMA significantly outperform their corresponding vanilla methods.

Method	Venue	ImgNet-R	ArxivQA	VizWiz	IconQA	CLEVR	Flickr30k	MFT (↑)	MFN (↑)	MAA (↑)	BWT (↑)
Zero-shot	–	16.27	53.73	38.39	19.20	20.63	41.88	–	31.68	–	–
Individual	–	91.67	90.83	57.87	78.43	76.63	61.72	–	76.19	–	–
LoRA-FT	ICLR'22	58.03	77.63	44.39	67.40	61.77	58.22	76.89	61.24	76.55	-18.78
OLoRA	EMNLP'23	77.50	78.07	44.50	63.13	64.73	58.16	76.01	64.35	78.02	-13.99
MoELoRA	NeurIPS'24	70.07	77.70	44.69	50.03	54.03	57.34	71.17	58.98	75.08	-14.63
CL-MoE	CVPR'25	66.33	77.00	44.78	51.87	53.53	57.42	71.46	58.49	74.19	-15.56
SEFE	ICML'25	80.83	78.00	47.01	69.63	65.83	57.92	75.98	66.54	78.76	-11.33
HiDE	ACL'25	84.03	90.73	44.43	58.93	41.37	54.25	69.96	62.29	77.32	-9.20
HiDE*	ACL'25	86.00	90.60	45.33	66.13	49.03	52.30	70.24	64.90	78.61	-5.34
+ PMA	**Ours**	84.77	93.83	50.35	70.67	53.07	54.55	72.50	67.87	80.57	-4.63
DISCO	ICCV'25	87.43	93.07	46.96	68.13	65.70	56.69	75.87	69.66	81.60	-7.45
DISCO*	ICCV'25	88.13	95.00	46.65	71.50	53.33	56.02	75.20	68.44	81.36	-6.76
+ PMA	**Ours**	88.00	95.50	52.66	74.50	69.77	58.83	**77.57**	**73.21**	**83.98**	**-4.36**

Table 2: Main results of the InternVL-Chat-7B model on the UCIT benchmark.

Method	Venue	ImgNet-R	ArxivQA	VizWiz	IconQA	CLEVR	Flickr30k	MFT (↑)	MFN (↑)	MAA (↑)	BWT (↑)
Zero-shot	–	21.10	63.20	40.59	24.70	21.20	44.67	–	35.91	–	–
Individual	–	95.40	92.70	62.53	82.87	83.80	58.82	–	79.35	–	–
LoRA-FT	ICLR'22	75.90	77.60	44.57	68.37	69.20	58.03	73.95	65.61	78.62	-8.34
OLoRA	EMNLP'23	86.37	93.73	44.13	68.10	64.53	56.19	72.36	68.84	81.10	-3.51
MoELoRA	NeurIPS'24	72.03	78.00	44.82	68.83	65.87	57.67	73.04	64.54	77.78	-8.51
CL-MoE	CVPR'25	74.23	78.77	44.77	53.90	72.73	58.23	71.39	63.77	77.44	-9.15
SEFE	ICML'25	85.93	76.90	47.42	69.77	63.37	58.19	73.55	66.93	79.44	-7.94
HiDE	ACL'25	90.00	93.77	50.49	71.03	61.10	55.08	76.29	70.25	82.97	-7.25
HiDE*	ACL'25	88.73	91.83	46.69	61.03	66.23	54.68	73.07	68.20	81.74	-4.88
+ PMA	**Ours**	91.70	94.87	51.66	71.00	69.20	56.88	76.91	72.55	84.28	-4.36
DISCO	ICCV'25	92.13	94.13	48.11	73.90	67.53	58.06	78.92	72.31	84.24	-7.93
DISCO*	ICCV'25	91.17	94.87	47.02	69.37	71.83	57.08	73.70	71.89	82.85	-1.81
+ PMA	**Ours**	93.37	95.07	59.90	74.97	82.83	58.40	**79.04**	**77.42**	**86.59**	**-1.61**

where PMA is integrated into two strong MCIT baselines, HiDE and DISCO.

5.1.1 Results on UCIT. As shown in Tables 1 and 2, incorporating PMA consistently improves performance across all evaluation metrics. On the LLaVA-1.5-7B backbone, HiDE+PMA achieves uniform gains across all four MCIT metrics, increasing MFT from 70.24 to 72.50, MFN from 64.90 to 67.87, and MAA from 78.61 to 80.57, while simultaneously mitigating forgetting (BWT: −5.34 → −4.63). When combined with DISCO, PMA further enhances MFT (75.20 → 77.57), MFN (68.44 → 73.21), and MAA (81.36 → 83.98), with a substantially less negative backward transfer (BWT: −6.76 → −4.36), yielding the best overall performance among all compared methods.

Similar trends are observed on InternVL-Chat-7B. HiDE+PMA improves MFT/MFN/MAA/BWT from 73.07/68.20/81.74/−4.88 to 76.91/72.55/84.28/−4.36, and DISCO+PMA achieves the SOTA overall results, reaching MFT/MFN/MAA/BWT of 79.04/77.42/86.59/−1.61.

5.1.2 Results on MLLM-DCL. As shown in Tables 3 and 4, PMA consistently improves MCIT performance on the MLLM-DCL benchmark. On LLaVA-1.5-7B, HiDE+PMA improves MFT/MFN/MAA from 61.77/56.04/62.30 to 62.67/57.38/63.49, while reducing forgetting (BWT: −5.73 → −5.29). When combined with DISCO, PMA further boosts MFT (64.61 → 66.28), MFN (59.24 → 62.57), and MAA (64.01 → 65.38), with a substantially less negative backward transfer (BWT: −5.37 → −3.71), achieving the best overall performance among all compared methods.

Similar trends are observed on InternVL-Chat-7B. HiDE+PMA improves MFT/MFN/MAA/BWT from 66.47/62.55/67.24/−3.92 to 68.60/65.05/69.34/−3.55, while DISCO+PMA attains SOTA overall performance, reaching MFT/MFN/MAA of 69.87/67.25/70.16 with the lowest forgetting (BWT: −2.62).

5.1.3 Analysis. Across both benchmarks and backbones, PMA consistently improves MFN and MAA while mitigating forgetting, without sacrificing MFT. These results demonstrate that explicitly addressing projector-level forgetting leads to more stable cross-modal

Table 3: Main results of the LLaVA-1.5-7B model on the MLLM-DCL benchmark.

Method	Venue	RS	Med	AD	Sci	Fin	MFT (↑)	MFN (↑)	MAA (↑)	BWT (↑)
Zero-shot	–	32.29	28.28	15.59	35.55	62.56	–	34.85	–	–
Individual	–	78.15	58.20	52.77	49.32	88.02	–	65.29	–	–
LoRA-FT	ICLR'22	69.65	41.59	25.43	40.88	87.45	64.98	53.00	61.13	-14.97
OLoRA	EMNLP'23	74.64	44.42	30.02	41.47	87.15	65.16	55.54	62.12	-12.03
MoELoRA	NeurIPS'24	77.54	41.85	27.62	40.13	86.75	64.94	54.78	61.76	-12.71
CL-MoE	CVPR'25	71.34	46.84	26.33	41.17	88.74	66.06	54.88	61.79	-13.97
SEFE	ICML'25	77.26	50.37	37.21	40.87	86.82	65.01	58.51	63.63	-8.13
HiDE	ACL'25	74.31	48.95	33.21	38.54	81.55	60.77	55.31	60.68	-6.82
HiDE*	ACL'25	74.68	50.37	34.14	40.14	80.88	61.77	56.04	62.30	-5.73
+ PMA	**Ours**	75.94	52.38	35.44	42.46	80.68	62.67	57.38	63.49	-5.29
DISCO	ICCV'25	76.49	44.48	44.84	46.61	89.22	64.78	60.33	63.93	-5.57
DISCO*	ICCV'25	73.61	44.86	47.92	44.83	84.96	64.61	59.24	64.01	-5.37
+ PMA	**Ours**	76.69	44.25	52.64	49.42	89.83	**66.28**	**62.57**	**65.38**	**-3.71**

Table 4: Main results of the InternVL-Chat-7B model on the MLLM-DCL benchmark.

Method	Venue	RS	Med	AD	Sci	Fin	MFT (↑)	MFN (↑)	MAA (↑)	BWT (↑)
Zero-shot	–	31.16	29.81	14.06	33.93	64.32	–	34.66	–	–
Individual	–	81.49	66.42	54.56	54.48	91.24	–	69.64	–	–
LoRA-FT	ICLR'22	69.93	52.17	33.04	42.67	91.07	69.06	57.78	65.22	-14.11
OLoRA	EMNLP'23	74.48	54.16	39.60	48.30	88.54	65.51	61.02	65.83	-5.62
MoELoRA	NeurIPS'24	69.90	52.08	33.17	42.19	90.58	68.83	57.58	65.97	-14.06
CL-MoE	CVPR'25	78.12	52.51	35.53	42.69	91.24	69.22	60.02	67.60	-11.51
SEFE	ICML'25	78.21	57.59	51.45	44.65	91.37	69.55	64.65	68.84	-6.12
HiDE	ACL'25	75.40	57.66	36.73	41.48	88.59	65.26	59.97	65.94	-6.61
HiDE*	ACL'25	78.54	57.64	37.68	48.45	90.46	66.47	62.55	67.24	-3.92
+ PMA	**Ours**	81.05	60.17	40.14	50.92	92.95	68.60	65.05	69.34	-3.55
DISCO	ICCV'25	77.90	47.50	49.13	49.37	90.92	68.55	62.96	67.81	-6.98
DISCO*	ICCV'25	77.70	50.19	53.41	50.30	90.67	68.69	64.45	68.16	-4.24
+ PMA	**Ours**	81.20	56.63	53.07	52.70	92.66	**69.87**	**67.25**	**70.16**	**-2.62**

alignment under continual instruction tuning. Importantly, the observed gains are orthogonal to backbone-level continual learning strategies such as HiDE and DISCO, indicating that PMA effectively complements existing MCIT methods by targeting an overlooked yet critical source of performance degradation.

5.2 Ablation Study

We conduct an ablation study based on the DISCO+PMA setting on the UCIT benchmark with the LLaVA-1.5-7B backbone to analyze the contribution of each component. As shown in Table 5, the full PMA configuration achieves the best performance across all four MCIT metrics. Removing the pretrained projector anchor ($\mathcal{P}^0$) preserves strong plasticity but results in noticeably worse MFN and backward transfer, indicating accumulated projector drift without a stable alignment reference. Disabling expert expansion causes performance to regress toward the original DISCO baseline, confirming that additional projector capacity is necessary to accommodate multimodal distribution shifts. Always expanding a new

Table 5: Ablation study on UCIT with LLaVA-1.5-7B under the DISCO+PMA setting. The bold denotes the highest result.

Method / Variant	MFT↑	MFN↑	MAA↑	BWT↑
DISCO* (shared projector)	75.20	68.44	81.36	-6.76
DISCO+PMA (full)	**77.57**	**73.21**	**83.98**	**-4.36**
w/o Anchor ($\mathcal{P}^0$ removed)	77.46	71.58	82.24	-5.88
No Expansion (single expert)	75.34	69.05	81.92	-6.29
Always Expand (one expert per task)	77.53	72.37	82.84	-5.16
Top-1 Routing (hard routing)	77.41	72.45	82.91	-4.96
Avg Weighting (uniform mixture)	77.18	71.91	82.43	-5.27

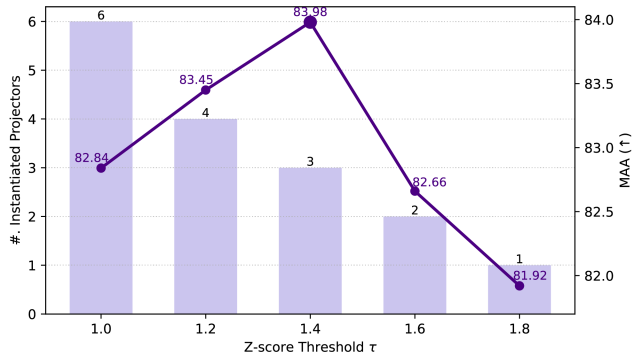


Figure 3: Effect of the z-score threshold τ on performance and projector growth under the DISCO+PMA setting on UCIT with LLaVA-1.5-7B. An intermediate threshold ($\tau = 1.4$) achieves the best MAA performance with sub-linear growth.

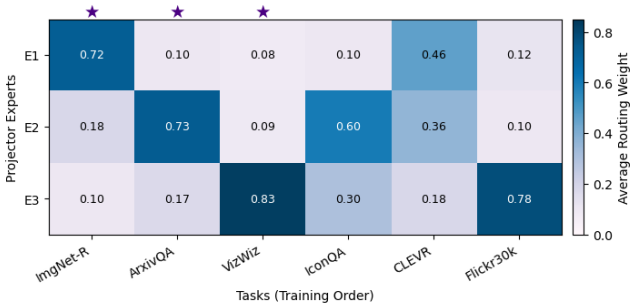


Figure 4: Projector usage analysis under the DISCO+PMA setting on UCIT with LLaVA-1.5-7B. Early tasks trigger projector expert expansion and mainly use their newly created experts, while later tasks reuse previously learned experts.

expert for each task slightly improves MFT but leads to degraded MFN and BWT, suggesting that uncontrolled expansion weakens cross-task sharing. Replacing the learned soft router with hard Top-1 routing or uniform averaging also yields inferior MFN and BWT, highlighting the importance of instance-wise, soft expert weighting. Overall, while all ablated variants outperform the original DISCO baseline by partially mitigating projector-level forgetting, only the full PMA design consistently achieves a strong balance between plasticity and stability.

5.3 Hyperparameter Analysis

We study the effect of the distribution-shift (z-score) threshold τ in PMA under the DISCO+PMA setting on UCIT with LLaVA-1.5-7B, focusing on the trade-off between continual performance and parameter growth. Figure 3 reports the resulting MAA and the number of instantiated projectors across different τ values. A smaller threshold (*e.g.*, τ=1.0) makes the detector overly sensitive, triggering expansions at nearly every task and leading to rapid parameter growth (6 experts), which weakens expert reuse and cross-task knowledge sharing, resulting in suboptimal performance (MAA=82.84). In contrast, a large threshold (*e.g.*, τ=1.8) rarely triggers expansion, yielding minimal growth (1 expert) but insufficient adaptation, degrading performance (MAA=81.92). Intermediate thresholds strike a better balance between efficiency and adaptability. In particular, τ=1.4 achieves the highest MAA (83.98) with only 3 experts, demonstrating that PMA attains strong MCIT performance with sub-linear projector growth by expanding capacity only when necessary.

5.4 Projector Usage Analysis

We analyze the projector expansion and reuse behavior of PMA under DISCO+PMA on UCIT with LLaVA-1.5-7B. After completing training on all 6 UCIT tasks, we evaluate the final model and visualize how PMA routes samples from each task's test set to different projector experts. Figure 4 reports the average routing weights of each task over the learned experts at the final evaluation stage. With τ=1.4, PMA instantiates 3 projector experts during training. The first three tasks (ImgNet-R, ArxivQA, and VizWiz), which trigger expert expansion, predominantly rely on the experts created for them, reflecting clear task-specific alignment needs as well as the bias toward newly added projectors described in Section 3.3. Importantly, later tasks exhibit substantial expert reuse. IconQA mainly reuses the expert trained on ArxivQA, while Flickr30k primarily reuses the expert trained on VizWiz, indicating that PMA captures transferable cross-modal alignment patterns among tasks with similar instruction styles and visual distributions. Notably, CLEVR shows mixed usage across multiple experts, which is consistent with PMA's soft reuse mechanism in Section 3.3, where the router assigns higher weights to the most compatible prior experts while allowing non-zero contributions from others. Since CLEVR combines short-answer outputs with structured reasoning, it benefits from integrating experts learned from both natural-image and structured QA tasks. Overall, these results show that PMA expands projector capacity only when necessary and achieves efficient sub-linear growth by flexibly reusing previously learned projectors.

6 Conclusion

In this paper, we identify projector-level forgetting as a critical yet largely overlooked challenge in MCIT. While existing approaches mainly focus on mitigating CF within the LLM backbone, we show that the shared projector responsible for cross-modal alignment can drift under sequential updates, leading to degraded instruction-following performance on previously learned tasks. To address this issue, we propose PMA, a method-agnostic framework that enables continual adaptation of the projector while preserving previously learned alignments. PMA detects multimodal distribution shifts using lightweight RDs, expands projector experts only when necessary, and integrates them via an expandable router anchored by the original pretrained projector. Extensive experiments on two MCIT benchmarks demonstrate that explicitly modeling projector-level adaptation consistently improves SOTA methods and scales effectively across different MLLM backbones.

Acknowledgment

This work was supported by the Brain Science and Brain-like Intelligence Technology - National Science and Technology Major

Project (2025ZD0217200), Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB1010302), CAS Project for Young Scientists in Basic Research (YSBR-116), Youth Innovation Promotion Association CAS, Shanghai Leading Talent Program of Eastern Talent Plan, the Lingang Laboratory Fund (Grant No. LG-GG-202402-06-07, LGL-1987-09), the Shanghai Municipal Science and Technology Project (Grant No. 25ZR1401370, 25LN3200400), Special Support Project of Guangdong Province (Grant No.0720240209). The numerical calculations in this study were carried out on the ORISE Supercomputer.

References

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

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923* (2025).

[3] Meng Cao, Yuyang Liu, Yingfei Liu, Tiancai Wang, Jiahua Dong, Henghui Ding, Xiangyu Zhang, Ian Reid, and Xiaodan Liang. 2024. Continual llava: Continual instruction tuning in large vision-language models. *arXiv preprint arXiv:2411.02564* (2024).

[4] Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. 2022. MapQA: A Dataset for Question Answering on Choropleth Maps. In *NeurIPS 2022 First Table Representation Workshop*.

[5] Cheng Chen, Junchen Zhu, Xu Luo, Heng T Shen, Jingkuan Song, and Lianli Gao. 2024. Coin: A benchmark of continual instruction tuning for multimodel large language models. *Advances in Neural Information Processing Systems* 37 (2024), 57817–57840.

[6] Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. 2023. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research* 20, 1 (2023), 38–56.

[7] Jinpeng Chen, Runmin Cong, Yuzhi Zhao, Hongzheng Yang, Guangneng Hu, Horace Ip, and Sam Kwong. 2025. SEFE: Superficial and Essential Forgetting Eliminator for Multimodal Continual Instruction Tuning. In *Forty-second International Conference on Machine Learning*.

[8] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*. Springer, 370–387.

[9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. 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*. 24185–24198.

[10] Jiahua Dong, Duzhen Zhang, Yang Cong, Wei Cong, Henghui Ding, and Dengxin Dai. 2023. Federated incremental semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3934–3943.

[11] Dong, Jiahua and Li, Hongliu and Cong, Yang and Sun, Gan and Zhang, Yulun and Van Gool, Luc. 2024. No One Left Behind: Real-World Federated Class-Incremental Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 4 (2024), 2054–2070. https://doi.org/10.1109/TPAMI.2023.3334213

[12] Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Jun Zhao, Wei Shen, Yuhao Zhou, Zhiheng Xi, Xiao Wang, Xiaoran Fan, et al. 2023. Loramoe: Revolutionizing mixture of experts for maintaining world knowledge in language model alignment. *arXiv preprint arXiv:2312.09979* 4, 7 (2023).

[13] Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, et al. 2025. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957* (2025).

[14] Chendi Ge, Xin Wang, Zeyang Zhang, Hong Chen, Jiapei Fan, Longtao Huang, Hui Xue, and Wenwu Zhu. 2025. Dynamic Mixture of Curriculum LoRA Experts for Continual Multimodal Instruction Tuning. In *Forty-second International Conference on Machine Learning*.

[15] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211* (2013).

[16] Haiyang Guo, Fanhu Zeng, Ziwei Xiang, Fei Zhu, Da-Han Wang, Xu-Yao Zhang, and Cheng-Lin Liu. 2025. HiDe-LLaVA: Hierarchical Decoupling for Continual Instruction Tuning of Multimodal Large Language Model. 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*. Association for Computational Linguistics, 13572–13586.

[17] Haiyang Guo, Fanhu Zeng, Fei Zhu, Wenzhuo Liu, Da-Han Wang, Jian Xu, Xu-Yao Zhang, and Cheng-Lin Liu. 2025. Federated continual instruction tuning. *ICCV* (2025).

[18] Haiyang Guo, Fei Zhu, Hongbo Zhao, Fanhu Zeng, Wenzhuo Liu, Shijie Ma, Da-Han Wang, and Xu-Yao Zhang. 2025. Mcitlib: Multimodal continual instruction tuning library and benchmark. *ICCV 2025@Workshop on Multimodal Continual Learning* (2025).

[19] Ziyu Guo, Renrui Zhang, Hao Chen, Jialin Gao, Dongzhi Jiang, Jiaze Wang, and Pheng-Ann Heng. 2025. SciVerse: Unveiling the Knowledge Comprehension and Visual Reasoning of LMMs on Multi-modal Scientific Problems. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*. Association for Computational Linguistics, 19683–19704.

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

[21] Jinghan He, Haiyun Guo, Ming Tang, and Jinqiao Wang. 2023. Continual instruction tuning for large multimodal models. *arXiv preprint arXiv:2311.16206* (2023).

[22] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286* (2020).

[23] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. 2021. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. *ICCV* (2021).

[24] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.

[25] Tianyu Huai, Jie Zhou, Xingjiao Wu, Qin Chen, Qingchun Bai, Ze Zhou, and Liang He. 2025. 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*. 19608–19617.

[26] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European conference on computer vision*. Springer, 235–251.

[27] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. 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*. 4999–5007.

[28] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* 114, 13 (2017), 3521–3526.

[29] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 3045–3059.

[30] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2025. LLaVA-OneVision: Easy Visual Task Transfer. *Transactions on Machine Learning Research* (2025).

[31] 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*. 19730–19742.

[32] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. 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)*. 14369–14387.

[33] Adam Dahlgren Lindström and Savitha Sam Abraham. 2022. CLEVR-Math: A Dataset for Compositional Language, Visual and Mathematical Reasoning. In *Proceedings of the 16th International Workshop on Neural-Symbolic Learning and Reasoning as part of the 2nd International Joint Conference on Learning & Reasoning (IJCLR 2022), Cumberland Lodge, Windsor Great Park, UK, September 28-30, 2022 (CEUR Workshop Proceedings, Vol. 3212)*. CEUR-WS.org, 155–170.

[34] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 26296–26306.

[35] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[37] Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. 2023. Moelora: An moe-based parameter efficient fine-tuning method for multi-task medical applications. *arXiv preprint arXiv:2310.18339*

(2023).
[38] Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. 2020. RSVQA: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing* 58, 12 (2020), 8555–8566.
[39] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. IconQA: A New Benchmark for Abstract Diagram Understanding and Visual Language Reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*.
[40] Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*. Vol. 24. Elsevier, 109–165.
[41] OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/
[42] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*. 2641–2649.
[43] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. 2024. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*. Springer, 256–274.
[44] Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuan-Jing Huang. 2023. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 10658–10671.
[45] Ziao Wang, Yuhang Li, Junda Wu, Jaehyeon Soon, and Xiaofeng Zhang. 2023. Finvis-gpt: A multimodal large language model for financial chart analysis. *arXiv preprint arXiv:2308.01430* (2023).
[46] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. 2023. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 2247–2256.
[47] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review* 11, 12 (2024), nwae403.
[48] Yahan Yu, Duzhen Zhang, Yong Ren, Xuanle Zhao, Xiuyi Chen, and Chenhui Chu. 2025. Progressive lora for multimodal continual instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2025*. 2779–2796.
[49] Fanhu Zeng, Fei Zhu, Haiyang Guo, Xu-Yao Zhang, and Cheng-Lin Liu. 2025. Modalprompt: Dual-modality guided prompt for continual learning of large multimodal models. *EMNLP* (2025).
[50] Duzhen Zhang, Yong Ren, Zhong-Zhi Li, Yahan Yu, Jiahua Dong, Chenxing Li, Zhilong Ji, and Jinfeng Bai. 2025. Enhancing Multimodal Continual Instruction Tuning with BranchLoRA. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 5743–5756.
[51] Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. In *Findings of the Association for Computational Linguistics ACL 2024*.
[52] Tielin Zhang, Xiang Cheng, Shuncheng Jia, Chengyu T Li, Mu-ming Poo, and Bo Xu. 2023. A brain-inspired algorithm that mitigates catastrophic forgetting of artificial and spiking neural networks with low computational cost. *Science Advances* 9, 34 (2023), eadi2947.
[53] Hongbo Zhao, Fei Zhu, Haiyang Guo, Meng Wang, Rundong Wang, Gaofeng Meng, and Zhaoxiang Zhang. 2025. Mllm-cl: Continual learning for multimodal large language models. *arXiv preprint arXiv:2506.05453* (2025).
[54] Junhao Zheng, Xidi Cai, Qiuke Li, Duzhen Zhang, ZhongZhi Li, Yingying Zhang, Le Song, and Qianli Ma. 2025. Lifelongagentbench: Evaluating llm agents as lifelong learners. *arXiv preprint arXiv:2505.11942* (2025).
[55] Junhao Zheng, Chengming Shi, Xidi Cai, Qiuke Li, Duzhen Zhang, Chenxing Li, Dong Yu, and Qianli Ma. 2026. Lifelong learning of large language model based agents: A roadmap. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2026).
[56] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592* (2023).