ModalPrompt: Towards Efficient Multimodal Continual Instruction Tuning with Dual-Modality Guided Prompt

Fanhu Zeng[1,2], **Fei Zhu**[3], **Haiyang Guo**[1], **Xu-Yao Zhang**[1,2*], **Cheng-Lin Liu**[1,2]
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems, CASIA
[2]School of Artificial Intelligence, UCAS [3]Centre for Artificial Intelligence and Robotics, HKISI-CAS
{zengfanhu2022, guohaiyang2023}@ia.ac.cn, zhfei2018@gmail.com, {xyz, liucl}@nlpr.ia.ac.cn

Abstract

Large Multimodal Models (LMMs) exhibit remarkable multi-tasking ability by learning mixed instruction datasets. However, novel tasks would be encountered sequentially in dynamic world, which urges for equipping LMMs with multimodal continual instruction learning (MCIT) ability especially for diverse and challenging generative tasks. Existing MCIT methods do not fully exploit the unique attribute of LMMs and often gain performance at the expense of efficiency. In this paper, we propose a novel prompt learning framework for MCIT to effectively alleviate forgetting of previous knowledge while managing computational complexity with natural image-text supervision. Concretely, we learn prompts for each task and exploit efficient prompt fusion for knowledge transfer and prompt selection for complexity management with dual-modality guidance. Extensive experiments demonstrate that our approach achieves substantial **+14.26%** performance gain on MCIT benchmarks with remarkable ×**1.42** inference speed free from growing computation. Code is available at `https://github.com/AuroraZengfh/ModalPrompt`.

1 Introduction

In recent years, large multimodal model (LMM), which aligns visual encoder (Dosovitskiy et al., 2021) with large language model (LLM) to handle multimodal tasks, has gained remarkable performance in numerous fields (Li et al., 2023; Liu et al., 2024b). As models become larger (Dubey et al., 2024; Zeng et al., 2025b,a), they are expected to perform lifelong learning like humans and learn more than one time to handle multiple tasks other than single tasks (Yao et al., 2022; Dai et al., 2024).

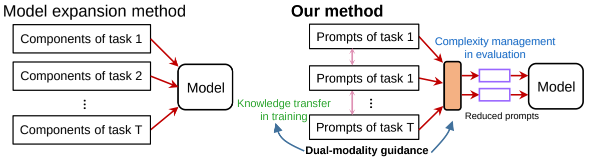

Figure 1: Diagram of model expansion method and our method. With natural attribute of multimodal guidance, we enhance MCIT with knowledge transfer and manage complexity against linear growth.

However, while pre-trained models like LLaVA perform well on mixed datasets, they tend to forget older tasks when fine-tuned on new task. Such catastrophic forgetting phenomenon is especially evident in sequential learning of widely differing multimodal tasks such as VQA and grounding (Goyal et al., 2017; Deng et al., 2021). This calls for multimodal continual instruction tuning (MCIT), which aims at sequentially fine-tuning models with multimodal instruction datasets and gets superior performance on new tasks while maintaining ability on previous tasks.

Existing approaches mainly tackle the forgetting issue by continually extending model with separate lightweight component for each task shown in Fig. 1, and LoRA (Hu et al., 2022) appears to be the common practice for large models (Wang et al., 2023). However, they expand model size and inference time in proportion to the number of tasks since they ensemble separated components of each task during inference. As the number of tasks increases, the cost of storage and inference become unbearable, particularly in LMMs and therefore hinder their practical deployments in real-world scenarios. Moreover, current methods derived from language models are not specially designed for LMMs (Hu et al., 2023; Razdaibiedina et al., 2023) without fully exploiting information from vision side and inevitably perform poorly on multimodal benchmarks. The mentioned shortcomings naturally raise an open question: *Can we establish an effective MCIT framework designed for LMMs refraining from growing computational expansion?*

In this paper, we investigate how to retain information of older tasks from multimodality (*i.e.*, im-

[*]Corresponding Author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 12126–12141
November 4-9, 2025 ©2025 Association for Computational Linguistics

age and text) to fully exploit LMMs and therefore improve the performance of multimodal continual instruction tuning efficiently. Generally speaking, given that the primary distinction between LLM and LMM lies in their utilization of image features, we establish a general prompt learning framework for multimodal continual instruction tuning with supervision from multimodality. **First**, we build a set of prompts for each task to represent task-specific knowledge and a lightweight projection layer is exploited to extract *prototype features* from task-specific prompts. When multimodal inputs come, the off-the-shelf text and visual encoders are used to obtain *multimodal features* representing multimodal distribution of current input. *Prototype* and *multimodal features* are then matched with similarity denoted as *dual-modality guidance*. **Second**, to enhance knowledge transfer, prompts that are most relevant to current task are obtained and fused through dual-modality guidance to promote the performance. **Third**, to address the problem of computational complexity, prompt selection mechanism from dual-modality guidance is developed to maintain inference efficiency.

Our method has two advantages: **(1)** guidance features after tokenization (text) and projection (image) naturally align multimodality information and are effortless to guide knowledge transfer and selection of LMMs; **(2)** computational complexity is in proportion to token numbers other than task numbers, and can therefore manage time consumption by selecting proper tokens. Extensive evaluation on MCIT benchmark across diverse multimodal tasks validates that our method substantially boosts performance on older tasks and mitigates forgetting with great training and inference efficiency. Our contributions are summarized as follows:

- We propose ModalPrompt, the first prompt learning framework for multimodal continual instruction tuning to mitigate forgetting with the advantage of multimodal supervision.
- We construct prompts to retain knowledge of specific tasks and exploit an effective dual-modality guided prompt fusion and selection technique to ensure MCIT ability while managing computational complexity.
- We conduct extensive experiments on continual instruction tuning benchmark, and the results substantially outperform existing methods (**+14.26%**) with great efficiency.

2 Related Work

Large Multimodal Models (LMMs) (Liu et al., 2024b,a; Ye et al., 2024), which combine vision representation with large language models (LLMs) (Alayrac et al., 2022; Touvron et al., 2023), have exhibited predominant function in numerous multimodal tasks (Liu et al., 2023; Lu et al., 2024). They typically contain a LLM decoder with stacks of transformers to decode embeddings. Usually, they first process image pixels with a CLIP image encoder, align features with a linear projector and then generate responses with concatenation of both image-text representations in an autoregressive way as LLMs do. As full fine-tuning is time-consuming and resource-intensive, efficient tuning is the common practice for instruction tuning of large models (Han et al., 2024). Methods for parameter efficient tuning are mainly three-fold: adapter learning (Zhang et al., 2021; Satapara and Srijith, 2024; Lee et al., 2024), prompt learning (Zhou et al., 2022) and LoRA (Hu et al., 2022). They update models with a lightweight module in the form of intra-block parallel connections, prefixes among input embeddings and low-rank decomposition, respectively. Specifically, multimodal instruction tuning (Wang et al., 2024b; Liu et al., 2025) has been a promising direction in promoting performance of multimodal models with both LoRA (Shen et al., 2024; Xu et al., 2024) and prompt learning (Khattak et al., 2023). As an orthogonal direction, we exploit techniques for mitigating catastrophic forgetting in multimodal foundation models.

Continual Instruction Tuning (Guo et al., 2025b) goes beyond instruction tuning that adapts large models to understand and align with human instructions. It primarily solves the problem of catastrophic forgetting (Zhai et al., 2023; Guo et al., 2025a,c) when one large foundation model sequentially learns multiple tasks through instruction tuning datasets. With the extensive development of LMM, much attention and effort have been paid to multimodal continual instruction tuning (Wu et al., 2024; Zhang et al., 2024). However, mainstream methods focus on transferring CIT methods from language tasks (Wang et al., 2024c,a, 2023) with no special design for visual features (He et al., 2023). Recently, a multimodal continual instruction tuning benchmark named CoIN (Chen et al., 2024) has been established and MoELoRA (Dou et al., 2023) is adopted to align previous instructions. However, it suffers from severe performance drop, highlight-

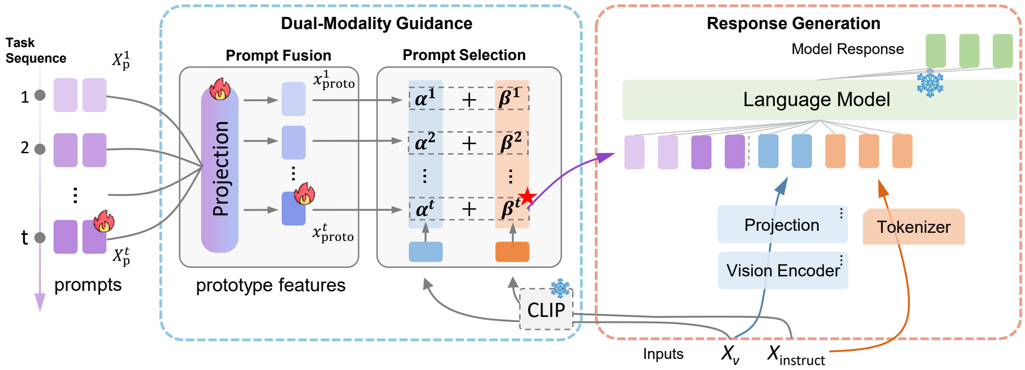

Figure 2: *Left:* prompt fusion module. Prototype features (x^t_{proto}) are obtained from the projection of prompts (X^t_{p}) to get task-specific knowledge in feature space. *Middle:* prompt selection process. Prototype features that are the most similar to current multimodal features are selected to enhance current input. *Right:* selected prompts and original multimodal inputs are concatenated and fed into large language model to generate responses.

Notation	Explanation
X^t_{p}	Prompts for task t
x^t_{proto}	Prototype features of prompts for task t
X_{v}, X_{insturct}	Image and text inputs of instruction tuning
x_{v}, x_{instruct}	Image and text features of guidance
α^t, β^t	Guidance from image and text modalities
$\widetilde{X}^t_{\text{p}}$, $\widetilde{X}^{\text{eval}}_{\text{p}}$	Selected prompts for training task t and evaluation

Table 1: Explanations of notations.

ing the necessity for exploration solutions tailored for multimodal continual instruction tuning.

3 Method: ModalPrompt

Notations. Multimodal continual instruction tuning seeks to address the issue of empowering LMM with ongoing datasets, where LMM $f_\theta(\cdot)$ is pre-trained on large-scale vision-language data to align image-text features. Given T tasks $\{\mathcal{T}^1, \cdots, \mathcal{T}^T\}$ with corresponding multimodal data $\mathcal{D}_t = \{X^{t,i}_{\text{v}}, X^{t,i}_{\text{instruct}}, y^{t,i}\}^{N_t}_{i=1}, t = 1, \cdots, T$, where $X^{t,i}_{\text{v}}$, $X^{t,i}_{\text{instruct}}$, $y^{t,i}$ stand for i^{th} sample of image, text and ground truth for t^{th} dataset (N_t in total), respectively. A continual learner aims to fine-tune $f_\theta(\cdot)$ sequentially on current data D_t while retaining knowledge on all previous tasks $\mathcal{T}^{<t}$ [1]. For clarification, Tab. 1 summarizes notations that would be used widely in this paper.

Problem setup. In this paper, we focus on multimodal continual instruction tuning in a more practical and challenging setting: (1) **Diverse generative tasks**: continual learning procedure is focused on generative tasks other than simple discriminative tasks like classification and with existence of vision information, task type is much more diverse with abundant scenarios; (2) **Free from task identification**: during inference, the model does not possess prior knowledge regarding which specific task current question belongs to; (3) **Absence of replay samples**: due to data privacy, no raw samples are replayed to refresh knowledge of previous tasks.

Overview. We present the basic prompt learning framework for multimodal continual instruction tuning. As illustrated in right of Fig. 2, the structure is similar to normal LMM, other than the input is prefixed with several prompts representing task-specific knowledge presented on the left. Given a set of prompts X^t_{p} with length M attached to each task $t, t \in \{1, \cdots, T\}$ representing task-specific knowledge in the form of MCIT, we focus on **(1) knowledge transfer** and **(2) complexity management**, which are **connected by dual-modality guidance**. Specifically, with dual-modality guidance matching multimodal input and task-specific knowledge, we propose multi-task prompt fusion to enhance knowledge transfer among different tasks during training in Sec. 3.2 and focus on how to manage complexity by prompt selection at inference time in Sec. 3.3.

3.1 Dual-Modality Feature as Guidance

The core for MCIT is how to bring **multimodal knowledge** of similar tasks to current input and generate response reasonably. Given that multimodal input naturally brings image-text supervision, we aim to match input and prompt in multimodal space. For image X_{v} and text X_{instruct} in each input of current task, considering that CLIP well captures image-text features, we reuse off-the-shelf vision and text encoder from CLIP to extract *multimodal features* of input:

$$x_{\text{v}} = \text{Proj}_{\text{v}}(E_I(X_{\text{v}})), \quad x_{\text{instruct}} = E_T(X_{\text{instruct}}), \quad (1)$$

where $E_I(\cdot) : \mathbb{R}^{\text{n}_{\text{v}} \times d_{\text{v}}} \rightarrow \mathbb{R}^{d_{\text{v}}}$, $E_T(\cdot) : \mathbb{R}^{\text{n}_{\text{t}} \times d_{\text{t}}} \rightarrow \mathbb{R}^{d_{\text{t}}}$ and $\text{Proj}_{\text{v}}(\cdot) : \mathbb{R}^{d_{\text{v}}} \rightarrow \mathbb{R}^{d_{\text{t}}}$ are CLIP vision

[1] we use the superscript for all elements from 1 to t.

encoder, text encoder and linear projection, respectively. $\mathrm{n_v}, \mathrm{n_t}, \mathrm{d_v}$ and $\mathrm{d_t}$ are length of image and text inputs, visual and textual dimension. The utilization can be effortless as they are well-trained and frozen for feature extraction, and the vision encoder has already been used in LMM to extract image features. We give detailed analysis of different encoders in Appendix C. The extracted text and visual features is crucial in enhancing continual ability (Sec. 3.2) and managing complexity (Sec. 3.3) described subsequently.

3.2 Training: Multi-Task Prompt Fusion

In contrast to class incremental learning (Zhu et al., 2021; Guo et al., 2024; Wang et al., 2022b) that learns distinct categories, datasets for MCIT like multiple types of question-answering tasks share general knowledge. Without explicit knowledge sharing among prompts of different tasks, in what follows, we first propose to transfer similar knowledge of older tasks in training procedure through multi-task prompt fusion to explicitly enhance MCIT. As illustrated in left of Fig. 2, prompts of all previous tasks are frozen for knowledge reuse and only current prompts are trainable, and we continually integrate knowledge of older tasks during sequential instruction tuning of current task with the aid of dual-modality features.

Prompt fusion for knowledge transfer. When training the t^{th} task, the trainable prompts are supposed to draw close to vision-language features of current task and absorb potential knowledge that may boost the performance. To enhance knowledge transfer, the dual-modality features could serve as guiding cues for prompts to accurately get close to multimodal distributions of current task in feature space. We propose to build prototype features from a lightweight projection layer (*e.g.*, MLP) to further align task-specific knowledge with guidance features from input:

$$x^t_{\text{proto}} = \text{Proj}_{\text{p}}(X^t_{\text{p}}), \tag{2}$$

where $\text{Proj}_{\text{p}}(\cdot)$: $\mathbb{R}^{M \times d_t} \rightarrow \mathbb{R}^{d_t}$ projects the prompts into task-specific *prototype features* in image-text feature space. It is effective in distinguishing whether prompts of older tasks are favorable for current tasks. Then, we explicitly match prompts and current input by fusing the prompts with the largest similarity of multimodal supervision for knowledge transfer. Concretely, we exploit the similarity between prototype features and dual-modality features as dual-modality guidance:

$$\alpha^t = \text{sim}(x^t_{\text{proto}}, x_{\text{v}}), \quad \beta^t = \text{sim}(x^t_{\text{proto}}, x_{\text{instruct}}), \tag{3}$$

where similarity is a measurement that matches current multimodal input and task-specific prompts and we exploit commonly used cosine similarity. With dual-modality guidance, the model has the ability to determine which prompts may boost the performance of evaluated task. We then select the prompts among $\{1, \cdots, t\}$ with k largest similarity of multimodal supervision:

$$\widetilde{X}^t_{\text{p}} = X^{\leq t}_{\text{p}} \circ \mathcal{I}_k\{\alpha^{\leq t} + \beta^{\leq t}\}, \tag{4}$$

where $\mathcal{I}_k : \mathbb{R}^{(M \times t) \times d_t} \rightarrow \mathbb{R}^{(M \times k) \times d_t}$ represents selecting the index with the largest k elements, and $\circ$ means selecting according to index. Note that in order to optimize parameters of current task, prompts of current task are always selected during training and prompts belonging to the same task are always selected simultaneously.

In summary, we explicitly integrate prompts that are close to the feature distribution of current task into training procedure by utilizing supervision from both modalities that caters for LMMs, and therefore transfer previous knowledge to boost the performance of current task.

Training objectives. During training, the inputs for continual learning of task $\mathcal{T}^t$ are prefixed with fused prompts $\widetilde{X}^t_{\text{p}}$ described in Eqn. 4. As shown in Fig. 2, parameters of large language model θ are frozen, and the introduced projection layer along with prompts corresponding to current task $\theta^t_{\text{p}} = \{\theta_{X^t_{\text{p}}}, \theta_{\text{Proj}_{\text{p}}}\}$ are trainable. The optimization target for task $\mathcal{T}^t$ is to find optimal parameters θ^t_{p} that minimize the negative log-likelihood language loss for LMMs:

$$\mathcal{L}^t_{\text{LMM}}(\theta^t_{\text{p}}) = \mathbb{E}_{(X^t_{\text{v}}, X^t_{\text{instruct}}, y^t) \sim \mathcal{D}_t} \Big[-\sum_{\ell=1}^{L} \log p(y^\ell | [\widetilde{X}^t_{\text{p}}, X_{\text{v}}, X_{\text{instruct}}, y^{<\ell}], \theta, \theta^1_p, \cdots, \theta^t_p) \Big]. \tag{5}$$

where L is the length of each sample pair in the dataset. The projection layer is optimized to reserve prototype feature during training process. Since we are to maximum the dual-modality guidance to keep knowledge of current task, we additionally design a prototype similarity loss to optimize prototype features formulated as:

$$\mathcal{L}^t_{\text{Proto}} = \Big[1 - \text{sim}(x^t_{\text{proto}}, x_{\text{v}})\Big] + \Big[1 - \text{sim}(x^t_{\text{proto}}, x_{\text{instruct}})\Big]. \tag{6}$$

Total training objective is a sum of language loss and prototype similarity loss:

$$\mathcal{L}_{\text{Total}}^{t} = \mathcal{L}_{\text{Proto}}^{t} + \mathcal{L}_{\text{LMM}}^{t}. \tag{7}$$

Parameters of current task are frozen afterwards to retain knowledge of learned tasks when learning new task.

3.3 Inference: Dual-Modality Prompt Selection

Another issue for MCIT is growing complexity with the number of tasks in evaluation. To handle this, we utilize dual-modality guidance for prompt selection in inference. As shown in middle of Fig. 2, by selecting the most relevant prompts, we convert the problem of computational complexity from task numbers $\mathcal{O}(T)$ to selected prompt numbers $\mathcal{O}(k)$, which greatly reduces computation load and improves efficiency.

Prompt selection for complexity management. After training on all sequential datasets, the crucial problem for evaluation is that the model has no ability to recognize which set of prompts promotes particular task and cannot manage computational complexity with simple ensembling leading to $\mathcal{O}(T)$ inference complexity. Intuitively, without access to data from older tasks, task-specific prompts should obtain cues for image-text distribution and be discriminant about which set of prompts counts during inference. To achieve this goal, we measure the similarity between image-text distribution of certain tasks and task-specific prompts employing dual-modality guidance. The representations of prototype features and multimodal feature are similar to Eqn. 2 and Eqn. 3, and differs in that selected prompts is determined among all T set of prompts:

$$\widetilde{X}_{\text{p}}^{\text{eval}} = X_{\text{p}}^{\leq T} \circ \mathcal{I}_k\{\alpha^{\leq T} + \beta^{\leq T}\}. \tag{8}$$

Intuitively, dual-modality guidance could serve as cues for selecting prompts that are helpful to current task in feature space, thereby converting computational complexity from $\mathcal{O}(T)$ to $\mathcal{O}(k)$.

Response generation. For each evaluated task, selected prompts $\widetilde{X}_{\text{p}}^{\text{eval}}$ with multimodal input $\{X_{\text{v}}, X_{\text{instruct}}\}$ are fed into LMM in a prefix way to generate answers:

$$f([\widetilde{X}_{\text{p}}^{\text{eval}}; X_{\text{v}}; X_{\text{instruct}}]; \theta), \tag{9}$$

where $[\cdot;\cdot]$ represents concatenation and $\widetilde{X}_{\text{p}}^{\text{eval}}$ is selected prompts through prompt selection.

Remarks. It can be concluded from above that the dual-modality guidance plays a crucial role in prompt learning for CIT and has two advantages: (1) help transfer knowledge from similar tasks to boost MCIT performance; (2) manage the inference speed as the time complexity is in proportion to the selected prompt numbers other than task numbers.

4 Experiments

We apply LLaVA (Liu et al., 2024b) as base LMM, and CLIP-Large-336 (Radford et al., 2021) as vision and text encoder for dual-modality feature extraction. Prompts can be easily constructed by extending the vocabulary size of language tokenizer. Length for each prompt M is set to 10. During prompt fusion and selection, we select 3 set of prompts. More implementation details can be found in Appendix B.

Datasets. We employ CoIN (Chen et al., 2024), a MCIT benchmark with numerous vision-language instruction datasets to evaluate continual instruction tuning ability. It includes OCRVQA (Mishra et al., 2019), GQA (Hudson and Manning, 2019), ImageNet (Deng et al., 2009), ScienceQA (Lu et al., 2022), Vizwiz (Gurari et al., 2018), TextVQA (Singh et al., 2019), VQAv2 (Goyal et al., 2017) and RefCoco (Mao et al., 2016; Kazemzadeh et al., 2014). Some of these datasets are visual question answering tasks of different fields, *e.g.*, GQA for visual reasoning and ScienceQA for science knowledge, and others are classification (ImageNet) and grounding (RefCoco). Following CoIN, we perform continual instruction tuning in the order of ScienceQA, TextVQA, ImageNet, GQA, VizWiz, REC, VQAV2 and OCRVQA and evaluate the performance after each continual stage.

Compared methods. Apart from MoELoRA (Dou et al., 2023), we implement three advanced prompt-based continual learning methods including L2P (Zhou et al., 2022), Dualprompt (Wang et al., 2022a) and CODA-Prompt (Smith et al., 2023) in the architecture of LMM for comprehensive comparison. We try our best to get optimal results and briefly introduce compared method and hyper-parameters in Appendix D.

Evaluation metrics. Denote that $A_{t,i} (i \leq t)$ is performance of task i after training on task t (T tasks in total). (1) For final performance evaluation, we measure each dataset using metrics ***Last*** (performance after sequential training on all tasks, *i.e.*, $A_{T,i}, i = 1, \cdots, T$) and ***Avg*** (average performance

	Method	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	REC	VQAV2	OCRVQA	*Average*
	Multi-task	81.43	61.36	90.05	60.67	52.39	66.14	63.54	61.28	67.10
	Zero-shot	67.92	57.71	47.37	61.28	45.17	6.12	52.03	53.58	48.89
Last	Finetune	26.00	25.38	28.51	33.07	26.52	0.10	40.00	52.92	29.06
	CODA-Prompt	58.15	50.16	24.04	54.33	48.94	17.83	55.86	54.42	45.46
	Dualprompt	56.40	47.12	34.96	42.03	44.14	12.01	54.43	53.36	43.05
	L2P	54.42	46.04	30.36	57.09	42.19	9.38	50.45	54.03	42.99
	MoELoRA	47.34	32.91	38.73	37.15	42.48	0.97	42.77	57.50	37.48
	Ours	**68.42**	**56.40**	**41.13**	**61.11**	**50.13**	**36.69**	**66.90**	**59.68**	**55.06 (+9.60)**
Avg	Finetune	13.79	15.74	9.08	28.84	15.20	0.06	40.00	-	17.53
	CODA-Prompt	48.84	47.17	18.74	50.77	42.68	15.43	55.86	-	39.93
	Dualprompt	42.81	43.41	24.12	40.52	40.39	12.76	54.43	-	36.92
	L2P	43.76	41.35	18.28	50.03	38.78	8.77	50.45	-	35.91
	MoELoRA	39.12	27.10	20.01	40.65	28.72	1.36	42.77	-	28.53
	Ours	**68.36**	**56.30**	**39.66**	**61.45**	**50.02**	**36.66**	**66.90**	-	**54.19 (+14.26)**

Table 2: Comprehensive comparison of multimodal continual instruction tuning ability. Performance is measured with accuracy.

Guidance Metrics	Image	Text	Dual	Δ
Last	51.95	50.39	55.06	+3.11
Avg	50.35	49.02	54.19	+3.84

Table 3: Effectiveness of guidance from multimodal supervision. Dual-modality similarity guidance achieves the best results.

across MCIT procedure). (2) For time-dependent continuous evaluation, we evaluate continuous metrics at each incremental stage across all trained datasets. The metrics include ***Backward Transfer (B)*** and ***Mean Accuracy (M)***. Zero-shot and multi-task are also reported to stand for the lower and upper bounds of the benchmark. Detailed explanations of these metrics are in Appendix A.

4.1 Main Results

Final continual performance. Results of MCIT benchmark are shown in Tab. 2. It can be concluded that: **(1)** Existing LoRA-based and prompt-based methods shows limited promotion in MCIT, highlighting the necessity of specific methods for LMMs. By contrast, our method achieves remarkable improvements with substantial **+9.60%** and **+14.26%** gain, respectively. Notably, the results after sequential tuning even against multi-task training, strongly demonstrating the effectiveness of the dual-modality guided prompt learning framework. **(2)** When learning different types of tasks, our approach undergoes moderate performance drop and still gets competitive results other than losing the ability to respond to the task (*e.g.*, the performance of MoELoRA drops to near zero when evaluated on Grounding), indicating the continual learning ability of the proposed method. **(3)** *Avg* of previous methods drop significantly compared with *Last*, and our method has almost no degradation, consistently achieving superior performance across the continuous tuning. More comparison results can be found in Appendix C.

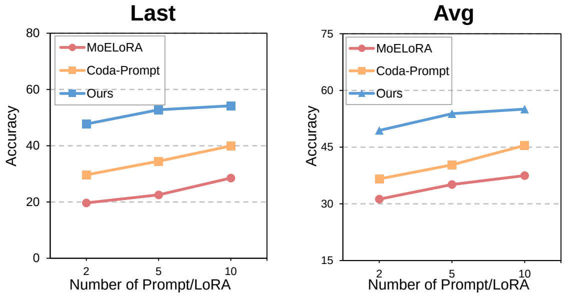

Figure 3: Impact of prompt/LoRA number. We implement different number of prompts for each task and different number of MoE for LoRA.

Continuous continual performance. We also evaluate continuous metrics at each incremental stage in Tab. 4 to examine time-variant multimodal continual instruction tuning performance. Concretely, compared with previous methods, our method is especially effective in alleviating catastrophic forgetting (BWT) to the most (**10.6%** mitigation) and also gets continuous promotion in across MCIT (**12.9%**). It is evident that our method outperforms state-of-the-art prompt-based method CODA-Prompt and LoRA-base method MoELoRA by a substantial margin with respect to both anti-forgetting and enhancing mean accuracy.

Method	TextVQA		ImageNet		GQA		VizWiz		REC		VQAV2		OCRVQA		*Average*	
	$B_2\downarrow$	$M_2\uparrow$	$B_3\downarrow$	$M_3\uparrow$	$B_4\downarrow$	$M_4\uparrow$	$B_5\downarrow$	$M_5\uparrow$	$B_6\downarrow$	$M_6\uparrow$	$B_7\downarrow$	$M_7\uparrow$	$B_8\downarrow$	$M_8\uparrow$	$B\downarrow$	$M\uparrow$
Finetune	44.30	44.14	65.53	32.52	64.42	22.75	51.98	25.55	67.08	5.71	37.62	23.64	31.16	29.06	51.73	26.19
CODA-Prompt	11.54	57.88	27.70	34.05	14.38	43.44	12.78	42.76	12.72	39.28	14.05	39.42	7.27	45.46	14.34	43.18
MoELoRA	41.31	43.13	52.47	34.08	32.76	41.71	33.81	37.71	41.41	25.59	30.80	34.34	26.12	37.48	36.95	36.29
Ours	**6.55**	**64.50**	**4.40**	**56.34**	**3.16**	**57.63**	**4.51**	**54.15**	**3.98**	**50.96**	**2.02**	**54.07**	**1.41**	**55.06**	**3.72 (-10.6)**	**56.10 (+12.9)**

Table 4: Continual performance metrics at each incremental stage. B_t and M_t stand for *Backward Transfer* and *Mean Accuracy* at incremental stage t.

Fusion	Selection	Last	Avg	B	M
	✓	37.36	31.87	27.09	34.94
✓		44.81	38.24	17.16	40.71
✓	✓	55.06	54.19	3.72	56.10

Table 5: Effectiveness of the proposed prompt selection and fusion for continual learning. Without prompt selection, we concatenate all prompts like Progressive Prompts (Razdaibiedina et al., 2023).

4.2 Ablation Study

We conduct numerous ablation studies to carefully validate the effectiveness of components and hyperparameters in the proposed method.

Effectiveness of dual-modality guidance. We develop the unique dual-modality guidance tailored for LMMs with multimodal information. To demonstrate the importance of multimodal guidance, we replace it with single-modality guidance. It is evident in Tab. 3 that either image or text information solely performs inferior to the proposed multimodal strategy, and vision information from multimodal dataset plays an inescapable function in guiding MCIT especially in datasets that rely heavily on image scenes. This strongly showcases that our method improves performance of MCIT by retaining robust and reliable prototype features in multimodal feature space and therefore contributes to all continuous tasks.

Prompt fusion and selection. We design the dual-modality guidance for knowledge transfer and complexity management, respectively. To validate the effectiveness of each proposed mechanism, we ablate each of them to demonstrate their usefulness. It is shown in Tab. 5 that both of them play a key role in the framework and lacking either of them causes severe performance drop. Specifically, multi-task prompt fusion is significant in promoting continual learning in the form of knowledge transfer. Besides, without selection, knowledge of different types of tasks would confuse the model and lead to performance drop. All results strongly highlight the effectiveness of dual-modality guidance in MCIT framework.

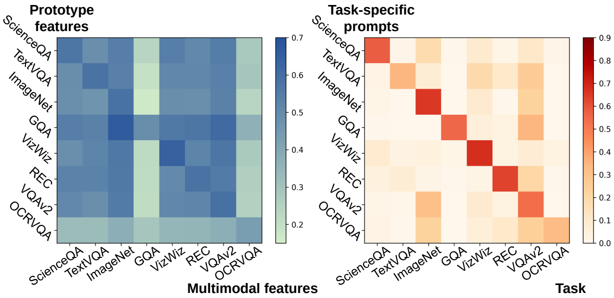

Figure 4: *Left:* Similarity between prototype features and multimodal task features. Larger value indicates more similar distribution. *Right:* Selection probability of each task from prompts. Results are percentage so the sum of rows equals one. Zoom in for better view.

Number of prompts. The number of prompts represents prototype features in aligned image-text space. We vary both numbers of prompt and LoRA in different methods to investigate the influence of prompt numbers. Results in Fig. 3 elucidate that increasing prompt numbers brings slight performance improvement. Considering both effectiveness and efficiency, we set the number of prompts for each task to 10 and do not further expand the quantity.

4.3 Further Analysis

Efficiency comparison. As prompt learning serves as another way to efficiently fine-tune large models, we compare our approach with LoRA-based (Chen et al., 2024) and prompt-based (Smith et al., 2023) method in terms of additional parameters, inference latency and GPU memory consumption to assess the efficiency. Tab. 6 reveals that our strategy achieves better results with lower memory, inference latency and trainable parameters. Specifically, we merely train **0.27%** of total parameters, which is **5%** of MoELoRA and **13%** of CODA-Prompt. Therefore, compared with baseline, our method achieves faster inference speed (×**1.42**), reduces training time (×**0.35**) and GPU memory consumption (×**0.92**), firmly substantiating the efficiency of our approach. The achievements can be attributed to simple prompt learning implementation and the prompt selection module that manages the computational complexity, consequently improving the

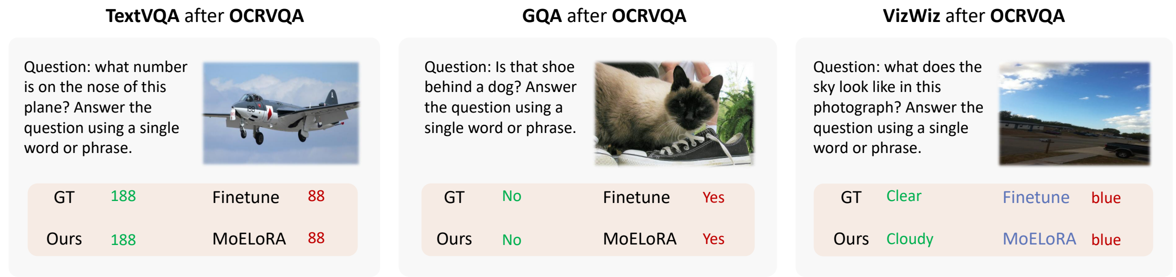
TextVQA after OCRVQA
Question: what number is on the nose of this plane? Answer the question using a single word or phrase.
GT 188
Finetune 88
Ours 188
MoELoRA 88
GQA after OCRVQA
Question: Is that shoe behind a dog? Answer the question using a single word or phrase.
GT No
Finetune Yes
Ours No
MoELoRA Yes
VizWiz after OCRVQA
Question: what does the sky look like in this photograph? Answer the question using a single word or phrase.
GT Clear
Finetune blue
Ours Cloudy
MoELoRA blue

Figure 5: Multimodal continual instruction tuning responses of several examples from TextVQA, GQA and VizWiz after fine-tuning on OCRVQA. Our method successfully mitigate forgetting and gives correct answers.

Method	Memory (M)	Training (Hour)	Trainable Param	Throughput (Token/s)
MoELoRA	16784	10.74	4.73%	2.41
CODA-Prompt	16073	5.12	1.97%	2.90
Ours	**15517**	**3.81**	**0.27%**	**3.43**
Δ	×**0.92**	×**0.35**	×**0.05**	×**1.42**

Table 6: Efficiency comparison of typical LoRA and prompt-based method with respect to GPU consumption, speed and trainable parameters. We average the training time for one epoch across datasets.

inference efficiency.

Similarity of dual-modality features. The ability of our framework to learn continually is largely guaranteed by the prompt selection module and prototype features represented in vision-language feature space. To further analyze the effectiveness of the dual-modality guidance tailored for LMMs, we calculate the similarity matrix between prototype features and multimodal task features. In Fig. 4, the similarity heatmap vividly illustrates the vision-language distributions of continual learning tasks. First, multimodal features of a few tasks are similar, indicating that most multimodal tasks share common sense and can promote each other mutually. However, some tasks, such as GQA and OCRVQA, are not similar to other tasks, which may be due to their task-specific ability not needed by other common tasks (visual reasoning for GQA and OCR for OCRVQA); second, the similarity is asymmetric, which may be attributed to their task inclusion relationship. For instance, GQA requires higher-level reasoning ability, while some other tasks may merely need to answer questions based on visual-language information. Therefore, features of GQA task (more basically) are similar to other tasks, but other tasks (more specifically) are not similar to the prototype of GQA. The visualization of dual-modality features exhibits the connection between prior obtained knowledge (prototype features) and given task (multimodal task features), and therefore contributes fundamentally to continual learning ability of LMMs.

Selection of prompts. To have an intuitive understanding of prompt selection module in addition to soft distribution construction, we report selection results of each previous task in percentage under MCIT setting to figure out the actual selection of prompts during inference. The results in right of Fig. 4 expose that the proposed module correctly matches and prioritizes prompts of the corresponding task as prefixes to enhance MCIT. Moreover, the module also selects prompts from similar type of tasks, which also enhances performance. This strongly indicates that knowledge transfer in tasks of the same type can mutually promote the performance, and our method leverages this characteristic excellently, demonstrating the robustness and usefulness of the learned prototype features.

Visualization. Fig. 5 provides examples during MCIT procedure to explicitly illustrate the effectiveness of our method. Specifically, existing methods fail on challenging multimodal generative tasks especially dependent on vision information. By contrast, our method can maintain performance on diverse previous tasks, keep knowledge from multimodality and answer challenging questions requiring comprehensive understanding correctly. For example, in TextVQA, the model identifies the specific part location of objects (nose of the plane) and overcomes occlusion; in GQA, it successfully distinguishes spatial orientation and therefore identifies objects. Moreover, it also deduces appropriate answers with analogous meanings to the ground truth based on image and text questions (*e.g.*, cloudy and clear in VizWiz). The visualizations demonstrate that based on the retained multimodal knowledge, our model gives the correct answer for diverse generative tasks, outperforming traditional existing continual instruction learning methods without design for vision information.

5 Conclusion

In this paper, we overcome the obstacle of continual learning tailored for LMMs with efficiency, and propose to exploit efficient prompt learning for continually learning image-text generative tasks while retaining knowledge of older tasks from multimodal supervision. Specifically, we construct a set of prompts for each task to represent task-specific knowledge in feature space. Building upon dual-modality guidance, we propose prompt fusion to enhance the performance from knowledge transfer and leverage prompt selection to manage the computational complexity of the model. Comprehensive experiments and analyses validate the effectiveness and efficiency of our framework.

Limitations

In this article, we propose ModalPrompt, an approach that exploits effective prompt fusion and selection with dual-modality guidance to retain performance in multimodal continual instruction tuning. While obtaining impressive continual learning performance, our method only retains knowledge of learned tasks and falls short of enhancing unseen tasks. However, we argue that it is an underexplored field as multimodal continual learning itself is not fully investigated. We will treat promoting forward transfer as future direction. Also, we believe that our model is generalizable and versatile, and plan to scale model size and application to other LMMs in future works.

Ethical Impact

We are committed to safeguarding intellectual property rights and complying with all applicable laws and regulations. The multimodal instructions included in our experiments are open-sourced from publicly available materials. We are dedicated to research purpose and are not intended for any commercial use.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62222609), National Science and Technology Major Project (2022ZD0116500), CAS Project for Young Scientists in Basic Research (YSBR-083), and the InnoHK program.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

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

Cheng Chen, Junchen Zhu, Xu Luo, Hengtao Shen, Lianli Gao, and Jingkuan Song. 2024. Coin: A benchmark of continual instruction tuning for multimodel large language model. *Advances in neural information processing systems*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. 2021. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Jun Zhao, Wei Shen, Yuhao Zhou, Zhiheng Xi, Xiao Wang, Xiaoran Fan, and 1 others. 2023. Loramoe: Revolutionizing mixture of experts for maintaining world knowledge in language model alignment. *arXiv preprint arXiv:2312.09979*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Haiyang Guo, Fanhu Zeng, Ziwei Xiang, Fei Zhu, Da-Han Wang, Xu-Yao Zhang, and Cheng-Lin Liu. 2025a. Hide-llava: Hierarchical decoupling for continual instruction tuning of multimodal large language model. *arXiv preprint arXiv:2503.12941*.

Haiyang Guo, Fanhu Zeng, Fei Zhu, Jiayi Wang, Xukai Wang, Jingang Zhou, Hongbo Zhao, Wenzhuo Liu, Shijie Ma, Da-Han Wang, and 1 others. 2025b. A comprehensive survey on continual learning in generative models. *arXiv preprint arXiv:2506.13045*.

Haiyang Guo, Fei Zhu, Wenzhuo Liu, Xu-Yao Zhang, and Cheng-Lin Liu. 2025c. Pilora: Prototype guided incremental lora for federated class-incremental learning. In *European Conference on Computer Vision*, pages 141–159. Springer.

Haiyang Guo, Fei Zhu, Fanhu Zeng, Bing Liu, and Xu-Yao Zhang. 2024. Desire: Dynamic knowledge consolidation for rehearsal-free continual learning. *arXiv preprint arXiv:2411.19154*.

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*, pages 3608–3617.

Zeyu Han, Chao Gao, Jinyang Liu, Sai Qian Zhang, and 1 others. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Zhiyuan Hu, Jiancheng Lyu, Dashan Gao, and Nuno Vasconcelos. 2023. Pop: Prompt of prompts for continual learning. *arXiv preprint arXiv:2306.08200*.

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

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Daehee Lee, Minjong Yoo, Woo Kyung Kim, Wonje Choi, and Honguk Woo. 2024. Incremental learning of retrievable skills for efficient continual task adaptation. *Advances in Neural Information Processing Systems*, 37:17286–17312.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yiyang Liu, James Chenhao Liang, Ruixiang Tang, Yugyung Lee, MAJID RABBANI, Sohail Dianat, Raghuveer Rao, Lifu Huang, Dongfang Liu, Qifan Wang, and 1 others. 2025. Re-imagining multimodal instruction tuning: A representation view. In *The Thirteenth International Conference on Learning Representations*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2023. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. 2023. Progressive prompts: Continual learning for language models. In *The Eleventh International Conference on Learning Representations*.

Shrey Satapara and PK Srijith. 2024. Tl-cl: Task and language incremental continual learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12123–12142.

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

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. 2023. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11909–11919.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Mingyang Wang, Heike Adel, Lukas Lange, Jannik Strötgen, and Hinrich Schütze. 2024a. Rehearsal-free modular and compositional continual learning for language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 469–480.

Taowen Wang, Yiyang Liu, James Liang, Junhan Zhao, Yiming Cui, Yuning Mao, Shaoliang Nie, Jiahao Liu, Fuli Feng, Zenglin Xu, and 1 others. 2024b. M2pt: Multimodal prompt tuning for zero-shot instruction learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3723–3740.

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*, pages 10658–10671.

Yifan Wang, Yafei Liu, Chufan Shi, Haoling Li, Chen Chen, Haonan Lu, and Yujiu Yang. 2024c. Inscl: A data-efficient continual learning paradigm for fine-tuning large language models with instructions. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 663–677.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and 1 others. 2022a. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, pages 631–648. Springer.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. 2022b. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 139–149.

Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. 2024. Continual learning for large language models: A survey. *arXiv preprint arXiv:2402.01364*.

Zhiyang Xu, Minqian Liu, Ying Shen, Joy Rimchala, Jiaxin Zhang, Qifan Wang, Yu Cheng, and Lifu Huang. 2024. Lateralization lora: Interleaved instruction tuning with modality-specialized adaptations. *arXiv preprint arXiv:2407.03604*.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2022. Filip: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051.

Fanhu Zeng, Zhen Cheng, Fei Zhu, Hongxin Wei, and Xu-Yao Zhang. 2025a. Local-prompt: Extensible local prompts for few-shot out-of-distribution detection. In *The Thirteenth International Conference on Learning Representations*.

Fanhu Zeng, Haiyang Guo, Fei Zhu, Li Shen, and Hao Tang. 2025b. Parameter efficient merging for multimodal large language models with complementary parameter adaptation.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2023. Investigating the catastrophic forgetting in multimodal large language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

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. *arXiv preprint arXiv:2401.13601*.

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

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

Didi Zhu, Zhongyi Sun, Zexi Li, Tao Shen, Ke Yan, Shouhong Ding, Kun Kuang, and Chao Wu. 2024. Model tailor: Mitigating catastrophic forgetting in multi-modal large language models. *arXiv preprint arXiv:2402.12048*.

Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. 2021. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5871–5880.

A Details of Evaluation Metrics

We give a thorough definition and explanation of the evaluation metrics used in the main experiments.

(1) Average: In addition to *Last*, which focuses on performance after tuning on all datasets, we propose to average performance throughout the entire tuning process. $\text{Avg}_i = \frac{1}{T-i}\sum_{t=i+1}^{T} A_{t,i}, i = 1, 2, \ldots, T-1$. It measures the absolute performance of each data across the sequential tuning. It is vital to keep the performance from dropping severely when the fine-tuning task varies greatly.

(2) Backward Transfer (BWT): It reflects the relative variation between current performance and direct tuning performance, measuring the catastrophic forgetting on all tasks. $B_t = \frac{1}{t-1}\sum_{i=1}^{t-1}(A_{i,i} - A_{t,i}), t = 2, \cdots, T$. Lower BWT represents better anti-catastrophic forgetting performance.

(3) Mean Accuracy (MA): $M_t = \frac{1}{t}\sum_{i=1}^{t} A_{t,i}$. It measures the average performance of all tasks at each incremental stage and is introduced to evaluate continual learning ability of all previous tasks. Higher MA stands for better continual learning ability. The above two metrics are averaged across all data on each incremental stage except the first one, *i.e.*, $t = 2, \ldots, T$.

B More Details of Experimental Settings

Continual instruction templates. For continual instruction tuning, the instructions for each datasets is shown in Tab. 7. Large language model concatenates instructions with image-text pairs in datasets to generate response accordingly.

Additional Implementation Details. Our framework is constructed depending on deepspeed repository [2] and Visual Instruction Tuning [3]. The instructions are from the repository of CoIN [4]. In evaluation of ImageNet, we give option choices for each question-answer pairs to avoid inaccurate descriptions. During training, all experiments are conducted on 48G NVIDIA A6000 and batch size is adaptively adjusted to maximize the memory utilization.

C More Experimental Results

Full continual instruction tuning results. We showcase brief results in the main results. We provide detailed continual instruction tuning performance during evaluation at each incremental stage. Upper, middle and bottom of Tab. 10 are full result comparison of different LMM continual instruction tuning approaches, including Finetune, MoELoRA and Ours. It can be concluded that our method achieves consistent and significant improvements against previous LoRA based method, validating the effectiveness of our method. Additional results of prompt-based methods are also shown in Tab. 11. It can be concluded that compared with prompt-based methods, our method also obtains substantial promotion, further certificating the utility of our approach.

Dataset	Instruction
ScienceQA	Answer with the option's letter from the given choices directly.
TextVQA	Answer the question using a single word or phrase.
ImageNet	What is the object in the image? Answer the question using a single word or phrase.
GQA	Answer the question using a single word or phrase.
VizWiz	Answer the question using a single word or phrase.
Grounding	Please provide the bounding box coordinate of the region this sentence describes: `<description>`.
VQAv2	Answer the question using a single word or phrase.
OCR-VQA	Answer the question using a single word or phrase.

Table 7: Instructions for each evaluated dataset.

Influence of different encoders. As stated in the paper, the dual-modality guidance coming from frozen CLIP-L plays a vital role in the proposed approach as the multimodal distribution from input is crucial to prompt selection and knowledge transfer. To explore the influence of different encoders, we replace text encoder with simple BPE tokenizer, and employ stronger vision encoder CLIP/G, respectively. Conclusion from Tab. 8 is that stronger encoder is slightly better. Yet, we conclude that **(1)** guidance merely selects but not represents knowledge and CLIP/L is fairly effective, which can be validated by Fig. 3 and Tab. 8; **(2)** as CLIP/L and image features have already been used in LLaVA, cost of computing and storage of CLIP-L is extremely small. Therefore, we use CLIP-L in current structure.

[2] https://github.com/microsoft/DeepSpeed

[3] https://github.com/haotian-liu/LLaVA

[4] https://github.com/zackschen/CoIN

	Vision Encoder	Text Encoder	ScienceQA	TextVQA	GQA	VizWiz	VQAv2	OCRVQA	Average
Last	CLIP/L	CLIP/L	67.82	56.41	60.76	51.08	66.93	59.52	60.42
	CLIP/L	BPE	62.42	53.73	62.27	45.92	65.11	62.49	58.66
	CLIP/G	CLIP/G	68.40	55.94	62.16	50.16	68.54	60.62	60.97
Avg	CLIP/L	CLIP/L	67.41	56.37	59.57	50.29	66.93	-	60.11
	CLIP/L	BPE	61.18	54.04	61.56	43.92	65.11	-	57.16
	CLIP/G	CLIP/G	68.29	56.11	62.28	49.97	68.54	-	61.04

Table 8: Influence of different encoders for continual instruction tuning on a subset.

Comparison with more methods. We compare our method with broader continual learning approaches including Model Tailor (Zhu et al., 2024), LWF (Li and Hoiem, 2017) and EWC (Kirkpatrick et al., 2017). Model Tailor maintains continual learning ability by calculating and enhancing important parameters for downstream tasks. It mainly focuses on reducing forgetting when fine-tuning a small number of downstream tasks and is not specifically designed for sequential adaptation across various tasks. As shown in Tab. 9, when encountering larger continual instruction tuning benchmarks, our method exhibits robust and better performance on the long continual learning process. Also, when compared with regularization-based approaches like LWF and EWC, our method obtains substantial improvements. These additional comparison comprehensively shows the advantage and effectiveness of our proposed method.

Method	**S**	**T**	**I**	**G**	**V**	**R**	**Q**	**O**	Avg
LWF	57.42	53.01	31.03	47.10	40.06	17.08	52.17	53.44	43.91
EWC	59.04	52.21	31.06	51.86	42.34	14.36	53.04	53.86	44.72
Model Tailor	**77.01**	44.09	26.33	47.28	37.16	25.40	54.06	56.73	46.01
Ours	68.42	**56.40**	**41.13**	**61.11**	**50.13**	**36.69**	**66.90**	**59.68**	**55.06**

Table 9: Additional comparison results with broader methods. **S**: **S**cienceQA, **T**: **T**extVQA, **I**: **I**mageNet, **G**: **G**QA, **V**: **V**izWiz, **R**: **R**EC, **Q**: V**Q**AV2, **O**: **O**CRVQA.

Experiment on other LMMs. We additionally conduct experiments on Qwen-VL (Bai et al., 2023), which integrates Q-former to align vision features with language models, to validate the effectiveness of the method on different LMM architectures. Consistent and substantial improvements in Fig. 6 certificate the adaptability and scalability of our method on different architectures, strongly demonstrating the extensibility and generalizability of the proposed method across different LMM architectures.

D Comparing Methods

We briefly introduce methodology of comparing approaches and then show the hyperparameters of each method. For practical implementation of existing methods, we have tried our best to get optimal results under fair comparison. Hyperparameters not mentioned are set by default.

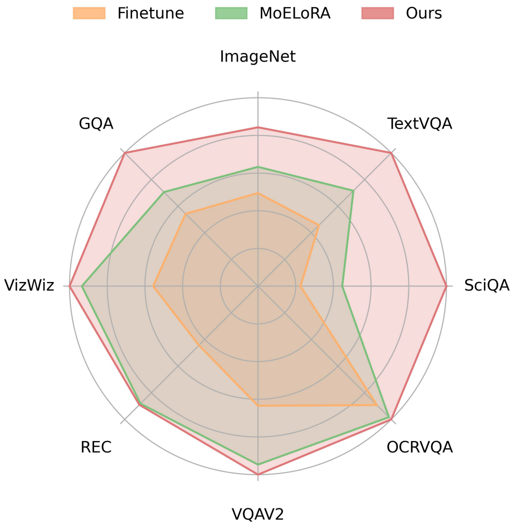

Figure 6: Results on Qwen-VL architecture.

MoELoRA leverages experts and gate function to activate part of parameters for each input. It learns knowledge of different tasks during training and mitigates forgetting in evaluation. We use 8 mixture of experts with rank $r = 16$.

L2P generates a pool of prompts in memory space. It manages task-invariant and task-specific knowledge in an explicit way of selecting relevant prompts for evaluation. We use pool size $M = 10$ with each length of each prompt $L_p = 10$ and select $N = 3$ for each task.

Dualprompt employs general prompt and expert prompt to encode task-invariant and task-specific instructions, and attach them to different layers of transformer block to meet the demand of knowledge restoration. We select number of general prompts $L_g = 3$ and number of expert prompts $L_e = 10$ practically.

CODA-Prompt proposes to learn a set of input-conditioned prompts for rehearsal-free continual learning. We use pool size $M = 10$ and length of each prompt $L_p = 10$.

Finetune	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	REC	VQAV2	OCRVQA
ScienceQA	82.45							
TextVQA	38.15	50.14						
ImageNet	0.96	0.58	96.03					
GQA	13.91	15.78	5.67	55.65				
VizWiz	8.46	25.17	4.60	38.12	51.42			
REC	0.00	0.00	0.00	0.27	0.00	34.00		
VQAV2	9.10	27.58	6.62	43.92	19.10	0.03	59.17	
OCRVQA	26.00	25.38	28.51	33.07	26.52	0.10	40.00	52.92

MoELoRA	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	REC	VQAV2	OCRVQA
ScienceQA	75.78							
TextVQA	34.47	51.80						
ImageNet	22.61	0.04	79.60					
GQA	32.37	34.04	42.48	57.95				
VizWiz	45.32	38.13	2.63	43.80	58.70			
REC	58.76	9.08	5.64	31.87	11.45	36.77		
VQAV2	33.01	48.42	10.61	49.78	32.23	1.75	64.58	
OCRVQA	47.34	32.91	38.73	37.15	42.48	0.97	42.77	57.50

ModalPrompt	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	REC	VQAV2	OCRVQA
ScienceQA	77.05							
TextVQA	70.50	58.50						
ImageNet	68.57	58.18	42.26					
GQA	68.82	56.08	43.43	62.17				
VizWiz	67.48	55.05	37.60	61.81	48.81			
REC	66.58	55.68	35.92	61.95	48.74	36.88		
VQAV2	68.12	56.43	40.22	60.92	51.19	36.63	64.99	
OCRVQA	68.42	56.40	41.13	61.11	50.13	36.69	66.90	59.68

Table 10: Detail continual instruction tuning results of Finetune, MoELoRA and our method.

L2P	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	REC	VQAV2	OCRVQA
ScienceQA	72.83							
TextVQA	68.07	57.16						
ImageNet	32.05	26.73	39.43					
GQA	47.53	46.02	18.03	60.47				
VizWiz	65.94	37.68	1.72	56.29	47.90			
REC	5.74	42.96	33.92	39.44	39.64	1.87		
VQAV2	32.57	48.65	7.41	47.32	34.52	8.17	59.40	
OCRVQA	54.42	46.04	30.36	57.09	42.19	9.38	50.45	54.03

Dualprompt	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	REC	VQAV2	OCRVQA
ScienceQA	67.16							
TextVQA	52.20	53.12						
ImageNet	28.49	24.77	46.40					
GQA	49.70	47.94	12.06	55.10				
VizWiz	57.88	51.17	21.34	48.03	51.62			
REC	18.27	39.64	29.77	44.06	35.97	30.82		
VQAV2	36.77	49.85	22.48	27.96	41.08	13.51	61.27	
OCRVQA	56.40	47.12	34.96	42.03	44.14	12.01	54.43	53.36

CODA-Prompt	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	REC	VQAV2	OCRVQA
ScienceQA	70.26							
TextVQA	58.72	57.05						
ImageNet	36.96	34.95	30.26					
GQA	50.78	53.52	10.12	59.35				
VizWiz	55.37	47.21	6.78	56.43	48.01			
REC	33.56	47.67	32.07	44.62	43.39	34.42		
VQAV2	48.34	49.54	20.72	47.72	35.73	13.03	60.87	
OCRVQA	58.15	50.16	24.04	54.33	48.94	17.83	55.86	54.42

Table 11: Detail continual instruction tuning results of prompt-based methods, inluding L2P, Dualprompt and CODA-Prompt.