SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •
Special Topic: Large Multimodal Models

December 2024, Vol. 67, Iss. 12, 220107:1–220107:18
https://doi.org/10.1007/s11432-024-4234-4

Modality-experts coordinated adaptation for large multimodal models

Yan ZHANG[1], Zhong JI[1,2*], Yanwei PANG[1,2], Jungong HAN[3] & Xuelong LI[4]

[1]*School of Electrical and Information Engineering, Tianjin Key Laboratory of Brain-Inspired Intelligence Technology, Tianjin University, Tianjin* 300072*, China;*
[2]*Shanghai Artificial Intelligence Laboratory, Shanghai* 200232*, China;*
[3]*Department of Automation, Tsinghua University, Beijing* 100084*, China;*
[4]*Institute of Artificial Intelligence (TeleAI), China Telecom Corporation Limited, Beijing* 100033*, China*

Received 8 May 2024/Revised 9 September 2024/Accepted 11 November 2024/Published online 13 December 2024

Abstract Driven by the expansion of foundation models and the increasing variety of downstream tasks, parameter-efficient fine-tuning (PEFT) methods have exhibited remarkable efficacy in the unimodal domain, effectively mitigating the consumption of computational resources. Although recent research has shifted attention to the multimodal domain and achieved efficient parametric adaptation of large multimodal models (LMMs) for downstream tasks, they still encounter two limitations: (1) low performance; (2) poor compatibility. This work proposes a modality-experts coordinated adaptation (ModeX) method for the multimodal domain, offering an effective, plug-and-play, and lightweight adaptation architecture for diverse LMMs. Specifically, ModeX adaptively coordinates different modality experts in terms of the types of network structure and input data. Besides, an effective coordinator equipped with a routing algorithm is developed for generating corresponding weights, which centers on leveraging the synergy among multimodal data. Extensive experiments on 15 multimodal downstream benchmarks and five LMMs demonstrate that ModeX is capable of seamlessly adapting to diverse LMMs, outperforms the state-of-the-art PEFT methods and even exhibits superior performance compared with full fine-tuning methods. Notably, on NLVR^2 task, ModeX achieves 84.06% accuracy with only 12.0M trainable parameters, outperforming the full fine-tuning by 1.63%. Moreover, our ModeX method demonstrates superior stability and offers higher training efficiency, both in terms of training parameters and training duration. Our source code has been released at https://github.com/zhangy0822/ModeX.

Keywords large multimodal model, multimodal learning, vision-language pretraining, parameter-efficient fine-tuning, adapter, modality expert

1 Introduction

While the impressive transferability of large multimodal models (LMMs) across diverse multimodal tasks highlights their potential, the ever-growing models and the expanding range of tasks render the conventional full fine-tuning (FFT) approach unfeasible, primarily due to the substantial computational and storage requirements.

To address these challenges, researchers have been actively exploring alternative methodologies. For instance, linear probe tailors lightweight heads for each downstream task, thereby reducing the scale of fine-tuning. Notably, recent parameter-efficient fine-tuning (PEFT) methods have demonstrated their remarkable ability to extend foundation models across diverse domains. However, the existing methods typically suffer from two major limitations: (1) How to significantly boost the performance so that it approaches or even exceeds FFT? (2) How to seamlessly and effectively adapt the lightweight module to the existing LMMs with high compatibility?

As shown in Figure 1, we illustrate the pipeline comparison between traditional FFT and the existing PEFT methods, and compare their respective advantages and disadvantages. Existing methods are either suitable for single modality, such as Prefix learning [1], Adapter [2], and LoRA [3] in natural language

* Corresponding author (email: jizhong@tju.edu.cn)

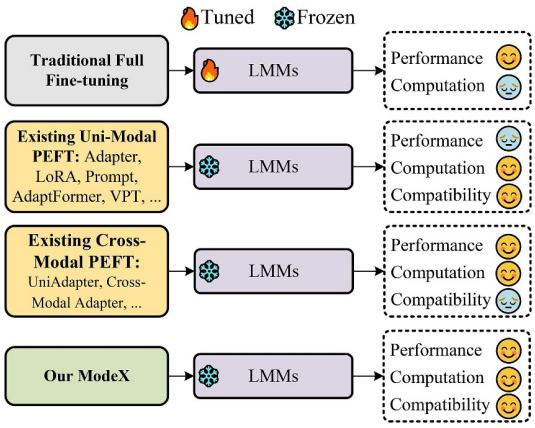

Figure 1 (Color online) Pipeline and applicability comparison among traditional FFT, existing PEFT methods, and our proposed ModeX method.

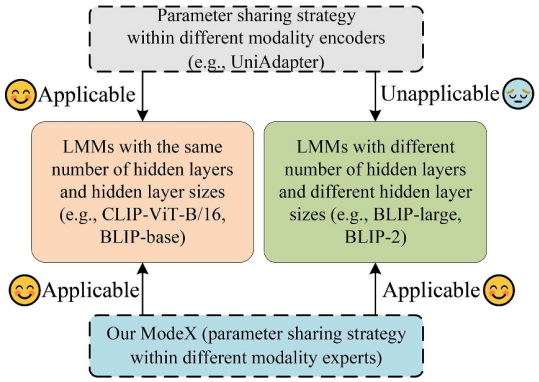

Figure 2 (Color online) Specific and intuitive explanation of the poor compatibility. The existing parameter sharing strategies within different modality encoders, especially Uni-Adapter [8] and Cross-Modal-Adapter [6], are unapplicable for most LMMs with different number of hidden layers and hidden layer sizes, e.g., BLIP-large, BLIP-2.

processing (NLP) and visual prompt tuning [4] in computer vision (CV), or they are task-specific and only suitable with limited LMMs, such as AIM [5] for video action recognition, Cross-Modal-Adapter [6] for video-text retrieval, and MultiWay-Adapter [7] for image-text retrieval. The above Uni-Modal methods may ineffectively adapt to LMMs and are incapable of performing as well as the tasks within their respective modalities (please refer to Subsection 4.2 for detailed results). In a nutshell, the first limitation is low performance. In the realm of multimodal, the Cross-Modal-Adapter [6], UniAdapter [8], and Aurora [9] introduce to share certain parameters, which exclusively applies if the image and text encoders have the same number of hidden layers and hidden layer sizes. The concurrent work [10] also reveals this limitation. As an illustration, both the number of hidden layers and hidden layer sizes of ViT-B/16 [11] and BERT-base [12] in BLIP-base [13] are 12 and 768, respectively. Nevertheless, most LMMs lack this nature, namely, image and text encoders do not share the same number of hidden layers or hidden layer sizes, as exemplified by BLIP-2 [14]. Specifically, the number of hidden layers and hidden layer sizes of ViT-L, querying transformer, and language model ($\text{OPT}_{2.7\text{B}}$ [15]) are 1024, 768, 2560 and 24, 12, 32 in BLIP-2, respectively. Thus, the second limitation is poor compatibility. An intuitive explanation is shown in Figure 2 [6, 8].

Faced with the above two limitations and the complex diversity of LMMs in function and structure, how to build an effective, plug-and-play, and lightweight parameter adaptation method? In LMM, different structures are responsible for implementing different functions, and their input modalities are also different. For example, the multi-head self-attention (MSA) layer in the ViT model exclusively supports visual embeddings as input and learns the output embeddings through intra-modality interaction; the Cross-Modal multi-head attention layer in BLIP supports multimodal embeddings as input and learns the output embeddings through inter-modality interaction. Thus, a unified taxonomy for different components of the LMMs is a crucial prerequisite. Besides, a pivotal insight lies in that different modalities of input data and different structures possess their respective experts with certain mutual synergy. We take BEIT-3 [16] with multiway transformer framework as an example, where a shared MSA layer encodes different modalities and a pool of feed-forward networks designed for different modalities.

To this end, we propose a modality-experts coordinated adaptation (ModeX) method for adapting various LMMs into downstream tasks, which coordinates the different modality experts in terms of different components. The main contributions of this paper are summarized as follows:

- We first introduce a unified taxonomy for most of the existing internal components of LMMs, which aims at establishing a criterion and improving the compatibility of the adaptation method.
- We propose a novel effective, plug-and-play, and lightweight parameter adaptation method, namely ModeX, which could be easily plugged to various LMMs and enables the efficient adapting to a spectrum of multimodal downstream tasks.
- In our ModeX method, we propose corresponding adaptation methods tailored to different types

of MSA and feedforward network (FFN) modules, designed to coordinate modality-specific experts and modality-mixture experts with a routing algorithm.

• Extensive experiments showcase that our ModeX method achieves comparable or higher performance and superior zero-shot transferability than previous FFT or PEFT methods on diverse multimodal tasks, e.g., 84.06% on NLVR^2 with 12.0M tunable parameters. Additionally, our ModeX method also demonstrates superior training efficiency in terms of both training parameters and training duration.

2 Related work

2.1 LMMs

Large-scale image-text pairs [17, 18] are collected and harnessed to engage in the pretraining of LMMs, which are dedicatedly trained for downstream tasks, such as image-text retrieval [19–23], image captioning [24, 25], visual question answering [26], and visual reasoning [27]. In terms of the downstream tasks they support, LMMs can be roughly divided into two-stream models [28, 29] and single-stream models [24, 30–32]. Recently, hybrid-stream models [13, 14, 33] support more multimodal tasks since they combine the advantages of the above two categories.

Despite the growing number of LMMs, their internal structure usually consists of the following parts: image encoder, text encoder, multimodal fusion encoder, and decoder. Further, this paper groups these components into Uni-Modal component and Cross-Modal component, and centers on developing versatile adaptation structure and efficiently adapting it into various LMMs.

2.2 PEFT

PEFT methods [1–3] initially emerged within the domain of NLP for alleviating the burdensome training and storage costs associated with the size of foundation models increases rapidly. Recently, the above PEFT and their various variant methods are applied in the domain of CV [4, 5, 34–36] and multimodal [6, 8, 37–39].

While pioneering studies have been set forth to explore efficiently fine-tuning, they suffer from low performance and poor compatibility. Specifically, Sung et al. [39] and Yang et al. [37] directly applied standard adapter or P-Tuning structure to LMMs, respectively, and did not get performance comparable to FFT. Despite Yang et al. [5] introduced a different usage of adapter, it focuses on a single modality and only adapts to video understanding task. Further, several studies [6, 8, 9] focused on the multimodal field and introduced efficient methods with different parameter sharing techniques. However, they are incapable of adapting to LMMs where encoders/decoders have different hidden layer sizes as well as hidden layer quantities. For instance, UniAdapter [8] and Cross-Modal-Adapter [6] introduce parameter sharing mechanisms between the visual encoder and textual encoder, which requires them to share the same number of hidden layers and the same hidden layer sizes. However, most LMMs are not equipped with these properties, such as BLIP-large [13], BLIP-2 [14], and LLaVA [40]. The mixture-of-modality adaptation (MMA) presented in [9] shares similarities with this study, particularly in the routing algorithm for weight allocation across different strategies. However, there are key distinctions in both function and structure between the two approaches. On the one hand, the MMA [9] is specifically developed to bridge the gap between large language models (LLMs) and vision-language tasks, and it only could be plugged at the beginning of each Transformer layer to process the multimodal data with explicit modality token. In contrast, our modality-experts coordinated adapter (MEC-Adapter) is designed to expand its compatibility for a wider range of tasks, and it could be plugged after the MSA layer and FFN layer or different combinations of them. On the other hand, while MMA [9] focuses on assigning routing weights exclusively for visual and text modalities, our approach introduces a novel parameter-sharing strategy and incorporates Cross-Modal experts to address the complexities of multimodal fusion and reasoning tasks (please refer to Subsection 4.2 for quantitative comparison results).

In this work, we center on developing an effective, plug-and-play, and lightweight module for various LMMs, which is capable of adapting to downstream tasks across a broad distribution. Specifically, we design an MEC-Adapter equipped with a novel parameter sharing strategy within different modality experts.

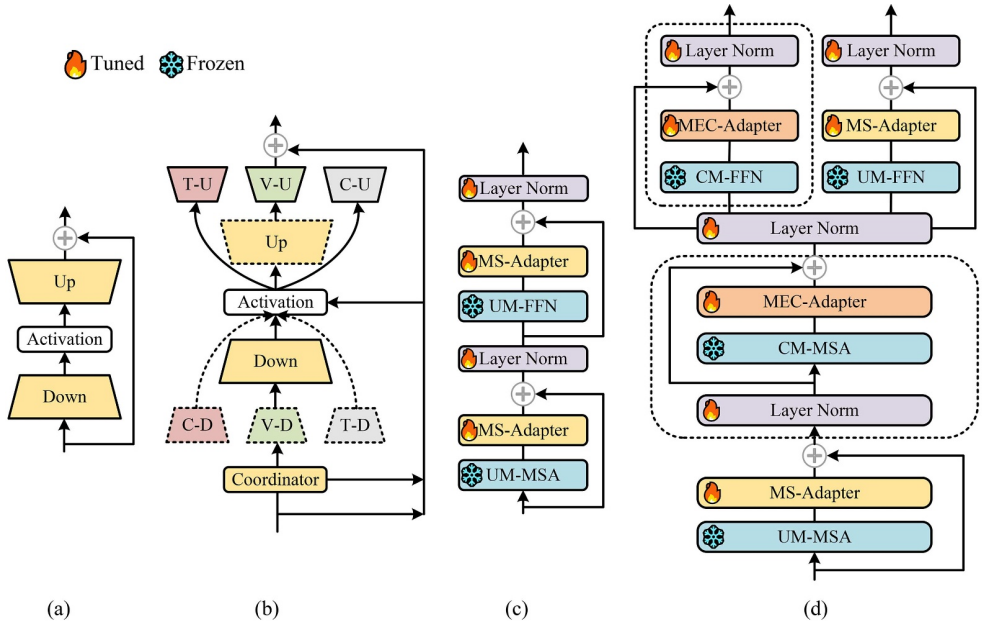

Figure 3 (Color online) Illustration of (a) modality-specific adapter (MS-Adapter), (b) MEC-Adapter, (c) visual encoder block, and (d) textual/multimodal encoder/decoder block. Note that the "UM" represents the "Uni-Modal" and "CM" represents the "Cross-Modal". Specifically, we show how we insert our proposed MS-Adapter and MEC-Adapter in (c) and (d), where the CM-MSA and CM-FFN in dashed boxes are alternative.

3 Methodology

In this section, we first introduce our unified taxonomy of LMM components in Subsection 3.1, which aims at providing criteria for our ModeX method and ensuring the compatibility of our method. Then, we propose our MEC-Adapter in Subsection 3.2. Finally, Subsection 3.3 presents how we adapt our proposed MEC-Adapter and MS-Adapter into LMMs for efficiently transferring to downstream tasks.

3.1 Preliminary

Unified taxonomy of LMM components. LMMs are equipped with varying network structures according to the types of tasks they support, such as encoder-only models [28,29] for retrieval tasks, encoder-decoder models [24, 41] for generation tasks, and hybrid-stream models [13, 14, 33] for understanding tasks. Despite they possess distinct network components, we explicitly divide the internal components of them into two categories: Uni-Modal/Cross-Modal multi-head self-attention (UM-MSA/CM-MSA) layer and Uni-Modal/Cross-Modal feedforward network (UM-FFN/CM-FFN). Thus, the underlying classification of LMM blocks comprises Uni-Modal encoder (including image and text) and Cross-Modal encoder/decoder with cross-attention layer.

Given the output of the $(l-1)$-th layer o_{l-1}, the detailed computing processings of Uni-Modal encoder are represented as follows:

$$o'_l = o_{l-1} + \text{UM-MSA}(\text{LN}(o_{l-1})), \tag{1}$$

$$o_l = o'_l + \text{UM-FFN}(\text{LN}(o'_l)), \tag{2}$$

where LN denotes the layer normalization operation. Similarly, given the output of the $(l-1)$-th layer o_{l-1} and the embeddings e from the other modality, the detailed computing processings of Cross-Modal encoder/decoder are represented as follows:

$$o'_l = o_{l-1} + \text{UM-MSA}(\text{LN}(o_{l-1})), \tag{3}$$

$$\hat{o}_l = o'_l + \text{CM-MSA}(\text{LN}(o'_l), \text{LN}(e)), \tag{4}$$

$$o_l = \hat{o}_l + \text{CM-FFN}(\text{LN}(\hat{o}_l)). \tag{5}$$

Adapter. Our method is grounded on the Adapter [2] framework, specifically developed for achieving parameter-efficient transfer learning within the domain of NLP. As shown in Figure 3(a), Adapter

encompasses a down-projection linear layer $\boldsymbol{W}_{\text{down}} \in \mathbb{R}^{(d\times r)}$, a non-linear activation function σ, and an up-projection linear layer $\boldsymbol{W}_{\text{up}} \in \mathbb{R}^{(r\times d)}$, where d denotes the dimensionality of the input and r denotes the bottleneck size. Then, the Adapter with identical architecture in each block is strategically inserted into the pre-trained model for adapting downstream tasks. Mathematically, given an input embedding $x \in \mathbb{R}^d$, the output of Adapter is formulated as follows:

$$\text{Adapter}(x) = x + \sigma(x\boldsymbol{W}_{\text{down}})\boldsymbol{W}_{\text{up}}. \tag{6}$$

3.2 ModeX

Distinct components handle different types of embeddings, such as the multi-head cross-attention layer and FFN in the multimodal encoder of ALBEF [33] (the detailed description and explanation are illustrated in Subsection 3.3). Thus, we focus on adaptively coordinating different modality experts in terms of the types of network structure and data, which aims at preserving and leveraging the information of different types of input embeddings.

As shown in Figure 3(b), we develop an MEC-Adapter, which can be regarded as a plug-and-play lightweight module and seamlessly inserted into different components within LMMs. In MEC-Adapter, we introduce our parameter sharing strategy, where similar strategies have been employed in existing methods [6, 42]. Differently, we design parameter sharing strategy on distinct modality experts within modules, such as MSA or FFN, instead of different encoders in [6, 8, 9]. Mathematically, given an input feature $x \in \mathbb{R}^d$ with multiple experts, MEC-Adapter captures different modality experts and adaptively coordinates them as follows:

$$\text{MEC-Adapter}(x, \text{UP}) = x + \sigma(x\boldsymbol{w}_{\text{down}}^{\text{share}}) \cdot \text{coordinator}(x), \tag{7}$$

where $\boldsymbol{w}_{\text{down}}^{\text{share}} \in \mathbb{R}^{(d\times r)}$ is the learnable weight matrix of down-projection linear layer, UP means that the coordination is deployed on up-projection. Similarly, we also provide an alternative scheme:

$$\text{MEC-Adapter}(x, \text{DOWN}) = x + \sigma(\text{coordinator}(x))\boldsymbol{w}_{\text{up}}^{\text{share}}, \tag{8}$$

where coordinator$(\cdot)$ denotes the specific architecture of our proposed coordinator, which is achieved by a simple routing algorithm:

$$\text{coordinator}(x) = \hat{r}^t x\boldsymbol{w}_{\text{up}}^t + \hat{r}^v x\boldsymbol{w}_{\text{up}}^v + \hat{r}^c x\boldsymbol{w}_{\text{up}}^c, \tag{9}$$

where $\boldsymbol{w}_{\text{up}}^t \in \mathbb{R}^{(d\times r)}$, $\boldsymbol{w}_{\text{up}}^v \in \mathbb{R}^{(d\times r)}$, and $\boldsymbol{w}_{\text{up}}^c \in \mathbb{R}^{(d\times r)}$ are three learnable weight matrices for adapting textual expert, visual expert, and Cross-Modal expert, respectively. $\hat{r}$ denotes the routing weights, which is computed as $\hat{r} = \text{softmax}(\frac{x\boldsymbol{w}_r + b_r}{\tau})$, where $\boldsymbol{w}_r \in \mathbb{R}^{(d\times 3)}$ and $b_r \in \mathbb{R}^3$ are the weight matrix and bias term, respectively. τ denotes the temperature of the softmax.

3.3 LMMs with ModeX

Most LMMs are built to achieve various multimodal tasks, such as image-text retrieval, image captioning, and visual reasoning. In Subsection 3.1, we represent our taxonomy criteria, where the internal components in LMMs are divided into UM/CM-MSA and UM/CM-FFN. To thoroughly explain our method and broaden the applicability, we choose ALBEF [33], BLIP [13], and BLIP-2 [14] as our frozen backbone from Salesforce as exemplars. More specific reasons are illustrated in Subsection 4.1.2.

We employ Figures 3(c) and (d) to illustrate the specific adaptation method, and the task-specific adaptation frameworks are delicately shown in Subsection 4.5. Note that we sequentially insert the MS-Adapter, and the rigorous experiments in Subsection 4.3 verify that it is more effective than parallel insertion [8, 43]. Particularly, for each image encoder block, two MS-Adapters after the UM-MSA and UM-FFN are inserted since only the visual embedding is input. Alternatively, the text encoder block has the same adaptation method as the image encoder block. Mathematically, given the output of the $(l-1)$-th layer o_{l-1}, the computing processings of Uni-Modal encoder with MS-Adapter are represented as follows:

$$o_l' = o_{l-1} + \text{MS-Adapter}(\text{UM-MSA}(\text{LN}(o_{l-1}))), \tag{10}$$

$$o_l = o_l' + \text{MS-Adapter}(\text{UM-FFN}(\text{LN}(o_l'))), \tag{11}$$

where o'_l and o_l denote the outputs of the l-th layer UM-MSA and UM-FFN, respectively. For each multimodal encoder/decoder block, we insert one MS-Adapter after UM-MSA, one MEC-Adapter after CM-MSA since the embeddings of another modality are introduced, and one MEC-Adapter after CM-FFN since the output encompasses the information of another modality. Mathematically, given the output of the $(l-1)$-th layer o_{l-1} and the embeddings e from the other modality, we formulate the computation of a standard multimodal encoder/decoder with MS-Adapter and MEC-Adapter as follows:

$$o'_l = o_{l-1} + \text{MS-Adapter}(\text{UM-MSA}(\text{LN}(o_{l-1}))), \tag{12}$$

$$\hat{o}_l = o'_l + \text{MEC-Adapter}(\text{CM-MSA}(\text{LN}(o'_l), \text{LN}(e))), \tag{13}$$

$$o_l = \hat{o}_l + \text{MEC-Adapter}(\text{CM-FFN}(\text{LN}(\hat{o}_l))), \tag{14}$$

where o'_l, $\hat{o}_l$, and o_l denote the outputs of the l-th layer UM-MSA, CM-MSA, CM-FFN, respectively.

4 Experiments

4.1 Datasets and settings

4.1.1 *Downstream tasks*

We divide the downstream tasks into two groups: (1) classic multimodal tasks; (2) emerging LLM-based zero-shot multimodal tasks.

(1) For the first group, we evaluate our proposed ModeX method on a wide range of vision-language tasks, including image-text retrieval: MSCOCO [44] and Flickr30K [45]; visual reasoning: NLVR^2 [27] and SNLI-VE [46]; visual question answering: VQAv2 [47]; and image captioning: MSCOCO (Karpathy test) [48] and NoCaps [49]. **Image-text retrieval.** Each image in MSCOCO [44] and Flickr30K [45] datasets is annotated with 5 captions. Following existing popular studies [8, 13, 14] and the Karpathy split [48], 113287/5000/5000 images on the MSCOCO dataset and 29000/1000/1000 on the Flickr30K dataset are employed for training/validation/testing, respectively. Following standard practices in Cross-Modal retrieval, we assess the retrieval performance on both caption retrieval (image query) and image retrieval (caption query) by R@K metrics ($K = 1, 5, 10$) and R@Mean. **Visual reasoning.** For NLVR^2, we conduct experiment on the official split [27], which contains 86k/7k/7k instances for training/validation/testing. This task requires the model to predict whether a sentence describes a pair of images. For SNLI-VE, we follow the original dataset split and obtain 29800/1000/1000 images for training/validation/testing. **VQA.** VQAv2 [47] dataset is adopted in this task, which contains 83k/41k/81k images for training/validation/testing. Following ALBEF [33] and BLIP [13], we employ both the training and validation splits for training, and include additional question-answer pairs from Visual Genome [50] dataset. During inference, we report the results of the test-dev split by submitting our predicted answers to the official server, while the results of the test-std split are not obtained due to the submission limitations of the official server. **Image captioning.** We fine-tune on the Karpathy train split of the MSCOCO dataset, and evaluate on the Karpathy test split of MSCOCO [48] and the NoCaps [49] validation set. Besides, we prepend a prompt "a picture of" at the beginning of each caption. During inference, we employ beam search with a beam size of 3, and set the maximum generation length as 20. For the NoCaps validation set, we report the results by submitting our generated captions to the official server.

(2) For the second group, the involved benchmarks including GQA [51], ScienceQA [52], MME [53], MM-Vet [54], POPE [55], Vizwiz [56], MMBench [57], and TextVQA [58]. Specifically, GQA [51] assesses the visual perception capability through open-ended short-answer questions. For ScienceQA [52], its image subset with multiple-choice is employed to assess the zero-shot generalization in scientific question answering. MME [53] measures both the perception and cognition abilities, offering a comprehensive assessment of models' multimodal understanding. MM-Vet [54] focuses on the integrated vision-language capabilities of LMMs. POPE [55] examines the hallucination levels across three subsets-random, common, and adversarial, respectively, and we report the F1 scores followed by [59]. Vizwiz [52] evaluates the zero-shot generalization on visual questions posed by visually impaired individuals. MMBench [57] offers a systematic and robust evaluation, designed to provide a holistic measure. As a classic optical character recognition (OCR)-related task, TextVQA [58] contains questions that could be answered by recognizing and reasoning about the text in images.

Table 1 Hyperparameters for PEFT on each downstream task

Task	Backbone	Optimizer	Learning rate	Schedule	Warmup (step, epoch, ratio)	Batch size	Epoch	Image resolution
Image-Text retrieval (MSCOCO [48], Flickr30K [45])	ALBEF [33]	AdamW	5e−4	Cosine decay	1000 steps	256	10	384
	BLIP [13]	AdamW	5e−4	Cosine decay	1000 steps	256	6	384
	BLIP-2 [14]	AdamW	5e−4	Cosine decay	1000 steps	224	5	364
	BEIT-3 [16]	AdamW	9e−3	Cosine decay	3 epochs	3072	15	384
NLVR^2 [27]	ALBEF [33]	AdamW	9e−4	Cosine decay	1000 steps	256	10	384
	BLIP [13]	AdamW	9e−4	Cosine decay	1000 steps	256	15	384
	BEIT-3 [16]	AdamW	5e−3	Cosine decay	5 epochs	256	20	224
SNLI-VE [46]	ALBEF [33]	AdamW	9e−4	Cosine decay	1000 steps	256	5	384
VQAv2 [47]	BLIP [13]	AdamW	5e−4	Cosine decay	3000 steps	256	10	480
Image captioning (MSCOCO [48], NoCaps [49])	BLIP [13]	AdamW	2e−4	Cosine decay	3000 steps	256	5	384
LLM-based zero-shot multimodal tasks	LLaVA-1.5 [59]	AdamW	5e−4	Cosine decay	0.03	128	1	336

4.1.2 *Implementation details*

As shown in Table 1 [13, 14, 16, 27, 33, 45–49, 59], we choose five popular LMMs with different network structures as our backbones to demonstrate the flexibility and compatibility of our ModeX, i.e., ALBEF [33], BLIP [13], BLIP-2 [14], BEIT-3 [16], and LLaVA-1.5 [59], which are pre-trained on millions of image-text pairs and the former three LMMs are uniformly built on the LAVIS [60] resrepository. There are three reasons for selecting the above five LMMs. (1) These five LMMs support a wide range of downstream tasks, including the classic multimodal tasks and the emerging LLM-based zero-shot multimodal tasks. (2) They possess different network structures and different levels of support for multimodal tasks, which proves the broader adaptability of our method. (3) They are advanced and representative LMMs in the field of multimodal. The parameters of these backbones are kept frozen during the fine-tuning process. We initialize the weights of down-projection layers for our proposed MEC-Adapter and MS-Adapter with Kaiming normal [61] and configure the weights of the up-projection layers with zero initialization. Most hyperparameters of our ModeX for each downstream task are listed in Table 1, and the temperature τ of our routing algorithm is 10. For the backbone of LLaVA-1.5 [59], we capitalize on the connector trained after the first feature alignment stage of the original paper [59] and directly carry out the second visual instruction tuning. The reasons are as follows: (1) the first feature alignment stage only involves a lightweight MLP connector without expensive training burden; (2) the same connector could ensure the fair comparison. Thus, we conduct the visual instruction tuning by injecting the 665K instruction-following data followed by [59], and evaluate the performance on the above LLM-based zero-shot multimodal tasks. All experiments are implemented with PyTorch and trained with two NVIDIA GeForce RTX 4090 GPUs.

4.2 Comparison with state-of-the-art methods

In this subsection, we compare our proposed ModeX with the existing state-of-the-art methods on a wide range of multimodal downstream tasks. We empirically categorize the existing methods into FFT and PEFT methods. The experimental results are shown in Tables 2–5, where the best and second-best results in each category are in bold and underlined. The symbol * denotes our reproduced results with a bottleneck dimension of 16. For a fair comparison, we fully fine-tune all the backbones under the same environment with our ModeX, which are denoted as ‡ in Tables 2–5.

Image-text retrieval. We evaluate our proposed ModeX for both image-to-text retrieval (TR) and text-to-image retrieval (IR) on MSCOCO and Flickr30K datasets, as shown in Table 2 [2, 3, 7, 8, 13, 14, 29–31, 33, 42, 43, 62–64]. For the backbone of BLIP, we observe that our ModeX achieves comparable results with FFT under the same backbone (the second-to-last lines of FFT), outperforms other PEFT methods while enjoying fewer trainable parameters. For the backbone of BLIP-2, we observe that our ModeX surpasses all the PEFT methods, and even outperforms the FFT method with fewer trainable parameters (87.2 R@Mean and 97.2 R@Mean with 5.9M trainable parameters on MSCOCO and Flickr30K, respectively).

Image captioning. As shown in Table 3 [2, 3, 13, 38, 39, 41–43, 65], we further verify our proposed ModeX for image captioning on MSCOCO Karpathy test [48] and NoCaps validation [49] datasets, where

Table 2 Comparison with the state-of-the-art methods for image-text retrieval on MSCOCO (5K) and Flickr30K (1K) datasets[a)]

Method	Parameters /model	Tunable parameters	MSCOCO TR			MSCOCO IR			MSCOCO	Flickr30K TR			Flickr30K IR			Flicker30K
			R@1	R@5	R@10	R@1	R@5	R@10	R@Mean	R@1	R@5	R@10	R@1	R@5	R@10	R@Mean
FFT methods:																
UNITER [30]	330M	330M	65.7	88.6	93.8	52.9	79.9	88.0	78.2	87.3	98.0	99.2	75.6	94.1	96.8	91.8
VILLA [62]	330M	330M	–	–	–	–	–	–	–	87.9	97.5	98.8	76.3	94.2	96.8	91.9
OSCAR [31]	330M	330M	73.5	92.2	96.0	57.5	82.8	89.8	82.0	–	–	–	–	–	–	
ALIGN [29]	820M	820M	77.0	93.5	96.9	59.9	83.3	89.8	83.4	95.3	99.8	**100.0**	84.9	97.4	98.6	96.0
ALBEF [33]	213M	213M	77.6	94.3	97.2	60.7	84.3	90.5	84.1	95.9	99.8	**100.0**	85.6	97.5	98.9	96.3
BLIP‡ [13]	223M	223M	81.4	95.0	97.8	64.0	85.9	91.6	86.0	96.5	99.6	99.7	87.5	97.8	99.0	96.7
BLIP-2‡ [14]	474M	474M	**82.8**	**96.0**	**98.2**	**66.6**	**87.1**	**92.4**	**87.2**	**96.9**	**100.0**	**100.0**	**88.7**	98.1	**99.2**	**97.2**
PEFT methods:																
P-tuning v2* [63]	BLIP	14.9M	77.5	93.7	96.9	61.0	84.0	90.1	83.9	91.8	99.3	99.6	80.7	94.9	97.3	93.9
Adapter* [2]	BLIP	1.5M	78.8	94.0	97.0	62.2	84.6	90.6	84.5	94.3	99.4	99.7	83.5	96.3	98.0	95.2
LoRA* [3]	BLIP	2.9M	79.4	94.2	97.0	62.7	84.8	90.8	84.8	95.8	99.7	99.8	84.4	96.8	98.3	95.8
AdaptFormer* [43]	BLIP	0.6M	77.1	93.0	96.4	61.0	83.7	89.8	83.5	92.7	99.2	99.5	80.5	95.3	97.3	94.1
Zero-Init Attention* [64]	BLIP	2.5M	74.5	92.4	96.0	60.1	83.1	89.6	82.6	91.2	99.2	99.8	80.0	94.5	96.9	93.6
MultiWay-Adapter [7]	BEIT-3	7.1M	78.3	–	–	60.7	–	–	–	95.4	–	–	85.4	–	–	–
UniAdapter [8]	BLIP	19.0M	80.1	94.6	97.4	62.6	84.6	90.9	85.0	97.1	**99.9**	**100.0**	86.4	97.4	98.9	96.6
LaVIN* [42]	LLaMA,CLIP-ViT	1.5M	79.8	94.9	97.6	63.1	85.2	91.1	85.3	96.9	99.8	**100.0**	87.0	97.4	98.9	96.7
ModeX (ours, $r = 16$)	BLIP	2.6M	80.4	95.1	97.7	63.7	85.3	91.3	85.6	96.9	**99.9**	**100.0**	86.8	97.5	98.8	96.7
ModeX (ours, $r = 32$)	BLIP	4.9M	80.4	95.3	97.7	63.7	85.6	91.3	85.7	**97.4**	**99.9**	**100.0**	87.1	97.6	98.9	96.8
ModeX (ours, $r = 16$)	BLIP-2	3.0M	83.1	95.8	**98.1**	66.7	86.9	92.2	87.1	96.8	99.7	**100.0**	**88.9**	**98.3**	**99.4**	**97.2**
ModeX (ours, $r = 32$)	BLIP-2	5.9M	**83.2**	**96.0**	98.0	**66.8**	**87.0**	**92.3**	**87.2**	96.7	99.8	**100.0**	**88.9**	**98.3**	99.2	**97.2**

a) Note that the results of BLIP-2 are fintuned on MSCOCO and zero-shot transferred to Flickr30K dataset.

Table 3 Comparison with the state-of-the-art methods on NoCaps validation and MSCOCO Karpathy test datasets[a)]

Method	Parameters /model	Tunable parameters	NoCaps validation								MSCOCO Karpathy test	
			In-domain		Near-domain		Out-domain		Overall			
			C	S	C	S	C	S	C	S	B@4	C
FFT methods:												
VL-T5 [41]	400M	400M	–	–	–	–	–	–	–	–	–	112.2
VinVL [65]	157M	157M	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5	38.2	129.3
BLIP‡ [13]	223M	223M	**113.3**	**15.1**	**109.7**	**14.9**	**110.8**	**14.2**	**110.4**	**14.8**	**39.9**	**133.8**
PEFT methods:												
VL-ADAPTER [39]	VL-T5	31.9M	–	–	–	–	–	–	–	–	–	111.8
VL-PET [38]	VL-T5	29.2M	–	–	–	–	–	–	–	–	–	121.7
Adapter* [2]	BLIP	1.5M	110.8	14.9	106.7	14.4	104.9	13.7	106.9	14.3	39.5	130.7
LoRA* [3]	BLIP	2.9M	109.3	14.8	108.1	14.6	105.0	14.0	107.6	14.5	39.0	131.3
AdaptFormer* [43]	BLIP	0.6M	108.5	14.7	106.4	14.4	104.4	13.7	106.3	14.3	39.3	130.2
LaVIN [42]	LLaMA,CLIP-ViT	5.4M	–	–	–	–	–	–	–	–	37.8	131.7
ModeX (ours, $r = 16$)	BLIP	2.3M	111.0	15.1	**108.7**	**14.7**	109.6	14.3	109.2	**14.7**	**39.6**	132.4
ModeX (ours, $r = 32$)	BLIP	4.4M	**112.8**	**15.3**	107.9	14.6	**111.1**	**14.4**	**109.3**	14.6	**39.6**	**132.5**

a) Note that C, S, and B@4 denote CIDEr, SPICE, and BLEU@4, respectively.

the latter is evaluated in a zero-shot manner, i.e., the model is trained by exclusively fine-tuning on the Karpathy train split of MSCOCO. Our ModeX achieves better performances than all the PEFT methods. For the results of MSCOCO, the slight improvement may stem from the process of fine-tuning capitalizes on the same objective as pretraining, i.e., language modeling loss. However, the improvement is even more significant on NoCaps, which proves that the model fine-tuned by ModeX has superior zero-shot transferability.

Natural language visual reasoning (NLVR^2). Unlike other multimodal tasks, NLVR^2 requires the model to predict whether a sentence describes a pair of images, and the existing PEFT methods neglect to report its results [8, 9] since it requires additional network design. Following BLIP [13], we employ two cross-attention layers to process the two input images and a merged layer to combine them, where the former is initialized from the same pre-trained weights and the latter is randomly initialized. Thus, we deploy two MEC-Adapters behind them with shared parameters. As shown in Table 4 [2, 3, 8, 9, 13, 30, 31, 33, 38, 39, 41, 43, 63, 64, 66], our ModeX method outperforms all the existing methods (FFT and PEFT methods).

Visual question answering (VQA). We evaluate our ModeX on the VQAv2 dataset to verify the understanding and generative capacity, which requires the model to predict an answer given an image and a question. As shown in Table 4, the significant improvement compared with other PEFT methods proves the effectiveness of our method on the VQA task.

Results of other backbones. In addition to the backbone of BLIP and BLIP-2, we also evaluate the

Table 4 Comparison with the state-of-the-art methods on NLVR^2 and VQAv2 datasets

Method	Parameters/model	Tunable parameters	NLVR^2		VQAv2
			Dev	Test-P	Test-dev
FFT methods:					
OSCAR [31]	330M	330M	78.07	78.36	73.16
LXMERT [66]	183M	183M	74.90	74.50	72.42
UNITER [30]	330M	330M	77.18	77.85	72.70
ALBEF [33]	266M	266M	**82.55**	**83.14**	75.84
VL-T5 [41]	400M	400M	–	74.30	–
BLIP‡ [13]	223M	223M	82.43	82.00	**76.78**
PEFT methods:					
VL-Adapter [39]	VL-T5	31.9M	–	72.70	–
VL-PET [38]	VL-T5	29.2M	–	73.42	–
P-tuning v2* [63]	BLIP	22.5M/29.6M	75.44	75.18	69.18
Adapter* [2]	BLIP	9.2M/2.4M	76.98	76.92	70.23
LoRA* [3]	BLIP	11.5M/4.9M	78.16	77.88	71.52
AdaptFormer* [43]	BLIP	8.3M/0.9M	75.22	75.05	68.08
Zero-Init Attention* [64]	BLIP	10.2M/3.8M	74.98	74.52	68.27
UniAdapter* [8]	BLIP	26.6M/4.8M	79.85	80.05	73.72
LaVIN* [9]	LLaMA,CLIP-ViT	8.9M/1.2M	83.27	82.68	72.07
ModeX (ours, $r = 16$)	BLIP	10.0M/4.0M	83.37	83.25	74.92
ModeX (ours, $r = 32$)	BLIP	12.0M/7.5M	**84.06**	**83.30**	**75.19**

Table 5 Comparison with the FFT on the backbones of ALBEF [33] and BEIT-3 [16]

Method	Parameters/model	Tunable parameters	NLVR^2		SNLI-VE		Flickr30K	
			Dev	Test-P	Val	Test	TR@1	IR@1
ALBEF‡ [33]	266M/213M/213M	266M/213M/213M	81.98	82.73	80.62	81.11	95.4	85.4
ModeX (ours)	ALBEF	5.6M/4.0M/3.4M	82.60	83.34	80.75	80.88	94.4	84.0
Method	Parameters/model	Tunable parameters	NLVR^2		MSCOCO		Flickr30K	
			Dev	Test-P	TR@1	IR@1	TR@1	IR@1
BEIT-3‡ [16]	226M/271M/222M	226M/271M/222M	82.8	84.0	76.3	58.9	95.4	84.6
ModeX (ours)	BEIT-3	8.3M/3.6M/3.6M	81.9	83.6	75.9	58.1	95.4	84.8

performance of ALBEF [33] on NLVR^2, SNLI-VE, and Flickr30K and BEIT-3 [16] on NLVR^2, MSCOCO, and Flickr30K, as shown in Table 5.

Results of LLM-based zero-shot multimodal tasks. As shown in Table 6 [28,59,67], we provide a comparison of three methods: FFT, LoRA, and our proposed ModeX. Methodologically, we align with the LoRA method described in [59]. Specifically, we fine-tune the ModeX module, which is inserted within the LLM, alongside the MLP connector, while freezing the parameters of the vision encoder and the original LLM. From the results, we observe that our ModeX attains the highest accuracy rates on the ScienceQA [52], MME [53], and MM-Vet [54] benchmarks, with just 156.8M trainable parameters. On the VisWiz [56], MMBench [57], and the adversarial split of POPE [55] benchmarks, our ModeX also secures the second-best results. Specifically, our ModeX method significantly outperforms LoRA by margins of 52.7 and 31.7 on the perception and cognition splits of MME [53], respectively, and it slightly surpasses the FFT method by 19.16 on the perception split of MME [53]. Additionally, compared with the FFT, our ModeX also improves it by 1.9 on the image subset of ScienceQA [52]. For the task of TextVQA [58], our ModeX also demonstrates comparable performance to the LoRA method with 46% of the trainable parameters. However, consistent with their backbone, i.e., the LLaVA-1.5 model, the scene text recognition and reasoning capabilities need to be further enhanced with more specific techniques to make them competent for more difficult OCR tasks, such as ChartQA [68] and DocVQA [69]. This defect is also mentioned and discussed in Section 5. Based on these analyses, we conclude that our ModeX could be efficiently and seamlessly integrated into the LLaVA framework.

Table 6 Comparison with the state-of-the-art methods on LLM-based zero-shot multimodal tasks, where the MLP connector, the Vicuna-7B [67], and the CLIP ViT-L/336px [28] collectively form the LLaVA model[a)]

Method	Tunable parameters	SQA-IMG	VisWiz	GQA	POPE			MME		MM-Vet	MMBench	TextVQA
					Rand	Pop	Adv	Per	Cog			
LLaVA-1.5-FFT [59]	7B	66.8	**50.0**	62.0	87.3	86.1	84.2	1510.7	–	**31.1**	64.3	**58.2**
LLaVA-1.5-LoRA [59]	340.8M	68.6	48.4	**63.0**	**87.8**	**86.8**	**84.8**	1477.6	265.4	30.9	**66.7**	57.5
LLaVA-1.5-ModeX (ours)	156.8M	**68.7**	49.4	61.4	87.2	85.9	84.5	**1530.3**	**297.1**	**31.1**	64.5	56.3

a) The bold and underlined fonts represent the best and second-best results, respectively.

Table 7 Comparison methods with different adapters and combination modes on Flickr30K and NLVR^2 tasks[a)]

Method	Mode	MS-Adapter	MEC-Adapter	BLIP-Flickr30K		BLIP-NLVR^2	
				TR@1	IR@1	Dev	Test-P
I	Partial	✓		96.1	86.4	82.34	82.24
II	Partial		✓	97.3	86.6	83.19	82.79
III	Entire	✓		**97.4**	87.2	82.87	82.99
IV	Entire		✓	97.2	87.0	82.57	82.65
V	Hybrid	✓	✓	97.3	**87.4**	82.69	82.46
VI	Hybrid*	✓	✓	**97.4**	**87.4**	**83.41**	**83.45**

a) The bold fonts represent the best results.

Table 8 Impact of the adapter position and quantity on MSCOCO and NLVR^2 tasks[a)]

Method	UM-MSA	UM-FFN	CM-MSA	CM-FFN	BLIP-Flickr30K		BLIP-2-MSCOCO		BLIP-NLVR^2	
					TR@1	IR@1	TR@1	IR@1	Dev	Test-P
I	✓				96.1	86.4	82.8	66.0	82.70	82.39
II	✓	✓			96.5	86.6	82.8	**66.5**	82.83	82.78
III	✓	✓	✓		97.1	86.8	**82.9**	66.3	82.58	82.90
IV	✓	✓	✓	✓	**97.4**	**87.4**	**82.9**	**66.5**	**83.37**	**83.20**

a) Note that the symbol "✓" denotes that the corresponding adapters are inserted after the module. The bold fonts represent the best results.

Table 9 Comparisons of different weight-sharing strategies for different experts of ModeX on MSCOCO and NLVR^2 tasks[a)]

Method	Tunable parameters	BLIP-MSCOCO		BLIP-NLVR^2	
		TR@1	IR@1	Dev	Test-P
w/o share	3.3M/10.6M	82.7	66.4	83.39	82.83
Share up, share down	2.7M/9.3M	82.8	66.5	82.87	82.99
Share down, coordinate up	3.0M/10.0M	82.9	66.5	83.37	83.25
Share up, coordinate down	3.0M/9.9M	**83.1**	**66.7**	**83.77**	**83.34**

a) The bold fonts represent the best results.

4.3 Ablation studies

In this subsection, we comprehensively provide ablation studies on our proposed ModeX, including the effectiveness of component and adaptation mode, the impact of the adapter position and quantity, different parameter sharing strategies of ModeX, and an investigation of the stability of ModeX. Unless otherwise stated, the coordination is deployed on up-projection.

4.3.1 *Effectiveness of components and adaptation mode*

As shown in Table 7, we delicately design several comparison methods with different adapters and combination modes. Specifically, "Partial" means only the corresponding modules (UM/CM-MSA/FFN) are equipped with the correct lightweight adapters (MS-Adapter or MEC-Adapter), "Entire" means all the modules are equipped with the corresponding and identical lightweight adapters, "Hybrid" means all the modules are equipped with the corresponding but distinct lightweight adapters with interleaved form, and "Hybrid*" means our method. The comparison between methods I and II serves to elucidate the efficacy of our proposed MEC-Adapter. Besides, the results drop slightly if we deploy distinct modules with identical adapters, which is concluded by comparing methods III and IV. Furthermore, the comparison between methods V and VI underscores the superior suitability of our unified taxonomy of LMM

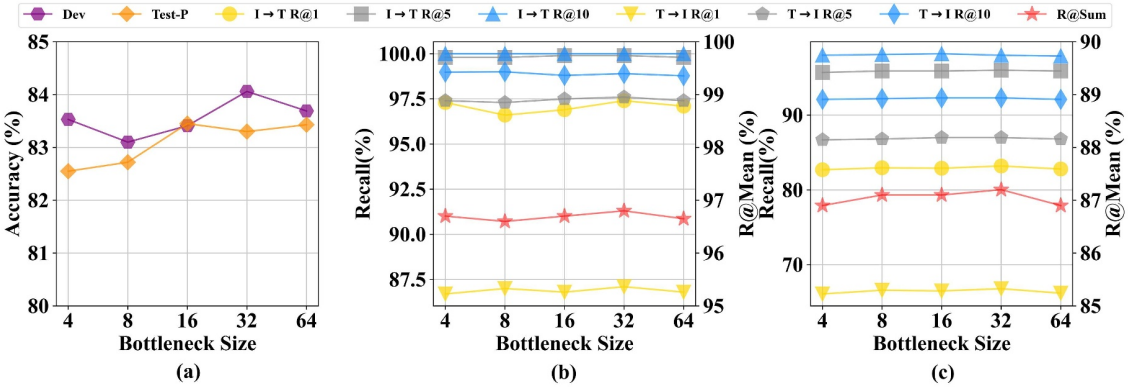

Figure 4 (Color online) Impact of bottleneck size of MS-Adapter/MEC-Adapter on the (a) BLIP-NLVR^2, (b) BLIP-Flickr30K (1K), and (c) BLIP-2-MSCOCO (5K) datasets. Note that R@K ($K = 1, 5, 10$) refer to the left vertical coordinates while R@Sum refers to the right vertical coordinates.

components and adaptation mode.

4.3.2 *Impact of the adapter position and quantity*

Just as the module categories introduced in Subsection 3.3, we deeply delve into the impact of the adapter position and quantity for the adaptation ability. As shown in Table 8, we report the performance of Flickr30K and NLVR^2 based on the BLIP model, and the performance of MSCOCO based on the BLIP-2 model, respectively. From methods I and II, as well as methods III and IV, we illustrate the effectiveness of the lightweight module after the FFN. The comparison between methods II and IV reveals the ability of our MEC-Adapter for coordinating the different expert information.

4.3.3 *Different parameter sharing strategies of ModeX*

Our proposed MEC-Adapter equipped with a simple yet highly effective parameter sharing strategy, that is, different experts share the same down-projection layer (up-projection) and set their respective up-projection layer (down-projection layer). As shown in Table 9, the reason that the parameters of the NLVR^2 task are generally higher than those of the MSCOCO retrieval task is that the former requires additional merge layers and classification layers, and the detailed frameworks are shown in Subsection 4.5. Besides, we compare four alternative strategies and observe that the "share up, coordinate down" achieves the optimal performance while enjoying preferable training efficiency. For multimodal tasks, the visual, textual, and Cross-Modal experts are initially extracted via three down-projections. They are then dedicatedly coordinated through shared up-projection, facilitating multimodal understanding and reasoning. Combined with the analyses of Table 9 and the LaVIN [42] results presented in Tables 2–4, we observe that the introduced parameter sharing strategy demonstrates higher effectiveness. The MEC-Adapter, which could be flexibly inserted after both the MSA and FFN layers, outperforms the previous approach that required insertion before the MSA layer with restrictions. Additionally, the effectiveness of the introduced Cross-Modal expert has been substantiated through these evaluations. Thus, we draw the conclusion that parameter sharing not only decreases the parameter dependence but also boosts the experts interaction for better performance.

4.3.4 *Investigation of the stability of ModeX*

The stability is also crucial in real scenario, whereas previous studies [8, 9] either overlook it or perform inferior. We empirically evaluate the stability of ModeX by observing the impact of different bottleneck sizes r, as shown in Figure 4. We observe the best r for R@Sum is 32, and the other metrics (R@1, R@5, R@10) also reach a relatively high level. Importantly, varying bottleneck size r does not cause much change in performance. Specifically, when r changes from 4 to 64, the accuracy on NLVR^2 is only from 82.55% to 83.45%, both higher than the results of FFT.

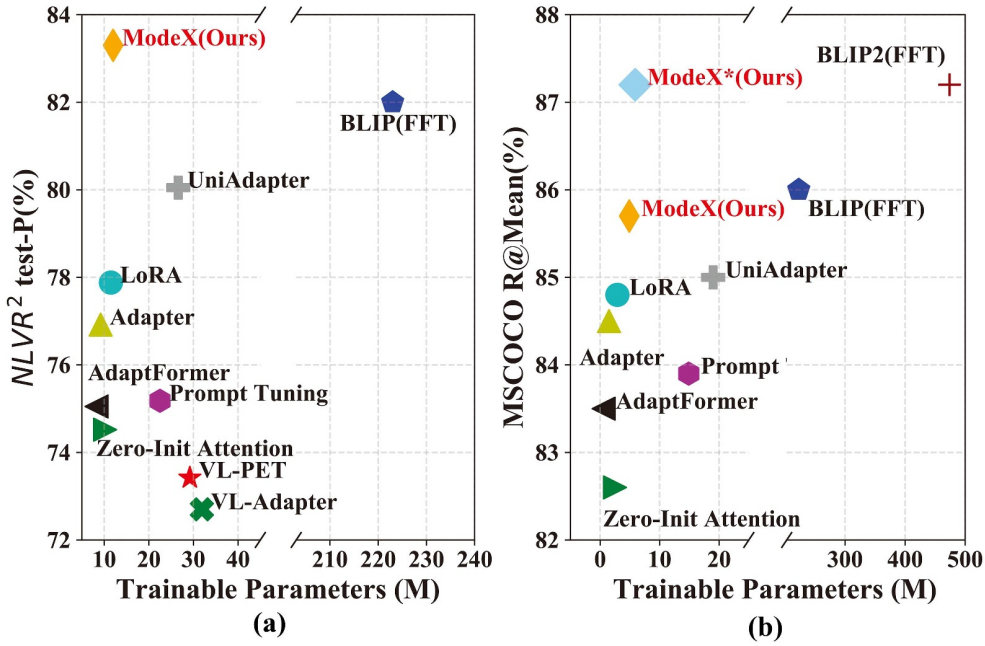

Figure 5 (Color online) Comparison of training parameters and performance trade-offs between ModeX and other methods on (a) NLVR^2 and (b) MSCOCO retrieval tasks, including the PEFT methods and two FFT models.

Table 10 Comparison of training duration in one epoch based on the backbone of BLIP [13]

Method	BLIP-Flickr30K		BLIP-MSCOCO		BLIP-NLVR^2		BLIP-Image captioning	
	R@Mean	Time (min)	R@Mean	Time (min)	Test-P	Time (min)	BLEU@4	Time (min)
P-tuning v2 [63]	93.9	50	83.9	390	75.18	39	129.6	122
Adapter [2]	95.2	42	84.5	218	76.92	28	130.7	91
LoRA [3]	95.8	54	84.8	385	77.88	36	131.3	116
ModeX (ours)	96.8	43	85.7	223	83.30	29	132.5	93

4.4 Trade-off between performance and training efficiency

Most of the existing PEFT methods only report the trainable parameters and neglect the importance of training duration in practice. Thus, we present these two types of testing in terms of training efficiency.

First, we show the comparisons of trade-off between performance and trainable parameters with the existing PEFT methods as well as FFT on two benchmarks. As shown in Figure 5, two subfigures illustrate the performance-trainable parameters trade-off, where the closer to the upper left corner, the better the overall performance. For the NLVR^2 task, the performance of our proposed ModeX method surpasses all the other PEFT methods, and the number of trainable parameters also achieves superior performance than most PEFT methods. Importantly, we observe that our ModeX method outperforms the FFT by 1.91% while enjoying extremely few parameters (12.0M vs. 223M). For the retrieval task, we compare two results of our proposed ModeX method, which are fine-tuned based on BLIP, BLIP-2 (ModeX*), respectively. From Figure 5(b), we conclude the following two observations: (a) the performance of our proposed ModeX method is the best among all PEFT methods, slightly below FFT; (b) ModeX* performs optimally among all the PEFT and FFT methods.

We further compare the trade-off between performance and training duration of different PEFT methods, as shown in Table 10. We observe that our ModeX requires nearly the same training duration as Adapter [2] in one epoch, and it outperforms those of P-Tuning V2 [63] and LoRA [3].

The above observations and conclusions illustrate that our proposed ModeX method performs superior than other methods, and achieves optimal trade-off between performance and training efficiency. More profoundly, our proposed ModeX method has stronger multimodal downstream task transferability than other PEFT and even FFT.

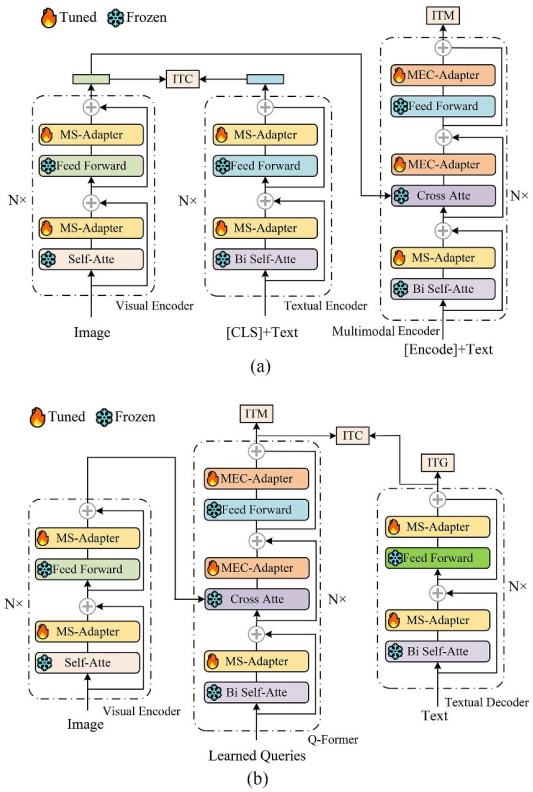

Figure 6 (Color online) Specific training frameworks on retrieval tasks based on (a) BLIP and (b) BLIP-2 models.

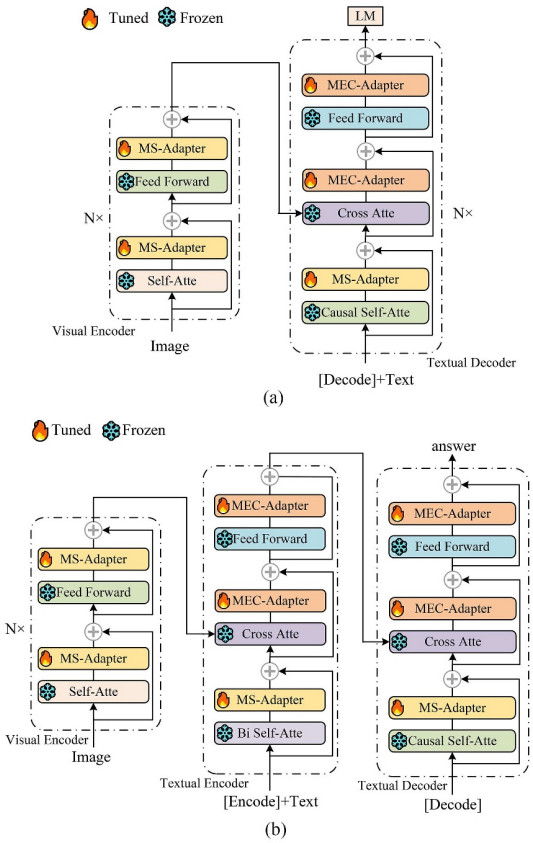

Figure 7 (Color online) Specific training frameworks on (a) image captioning and (b) VQA based on the BLIP model.

4.5 Case studies of model adaptation

In this subsection, we intuitively provide several specific model adaptation framework, including retrieval, reasoning, and generation tasks, which aims at exhibiting how to adapt our proposed adapters into an existing LMM in conjunction with specific downstream tasks, as shown in Figures 6–8. Note that layer normalization layers are omitted for simplicity, and "Atte", "ITM", "ITC", "ITG" denote the "Attention", "Image-Text Matching", "Image-Text Contrastive", and "Image-Grounded Text Generation" losses, respectively.

For the retrieval task, two training frameworks based on BLIP and BLIP-2 models are shown in Figure 6. Although these two frameworks have some of the same training objectives, such as ITM and ITC, their training and inference processes are quite different. Specifically, BLIP-2 inserts a cross-attention layer only for every two layers of querying Transformer blocks, and trains the framework with additional ITG objective.

For the generation task, we provide the frameworks of image captioning and VQA in Figures 7(a) and (b), where the above bidirectional self-attention layer is replaced with the causal self-attention layer. Both the tasks were fine-tuned with the LM objective. Differently in VQA, i.e., Figure 7(b), image and its question are first encoded into multimodal embeddings, which are input to an answer decoder.

For the reasoning task, especially the NLVR^2, which involves two images and a text, the model is asked to answer whether the text description is true or false. A shown in Figure 8, we follow the module construction method of BLIP and insert our proposed adapters into the corresponding positions. Specifically, two cross-attention layers initialized from the same pre-trained weights are built to process the two input images, and their outputs are performed average pooling in the first 6 layers and concatenated with a linear projection in the last 6 layers. Besides, an MLP module as additional classifier is adopted on the output of the [Encode] token. Notably, the merge layers and the MLP module require training because they are not trained during the pretraining process.

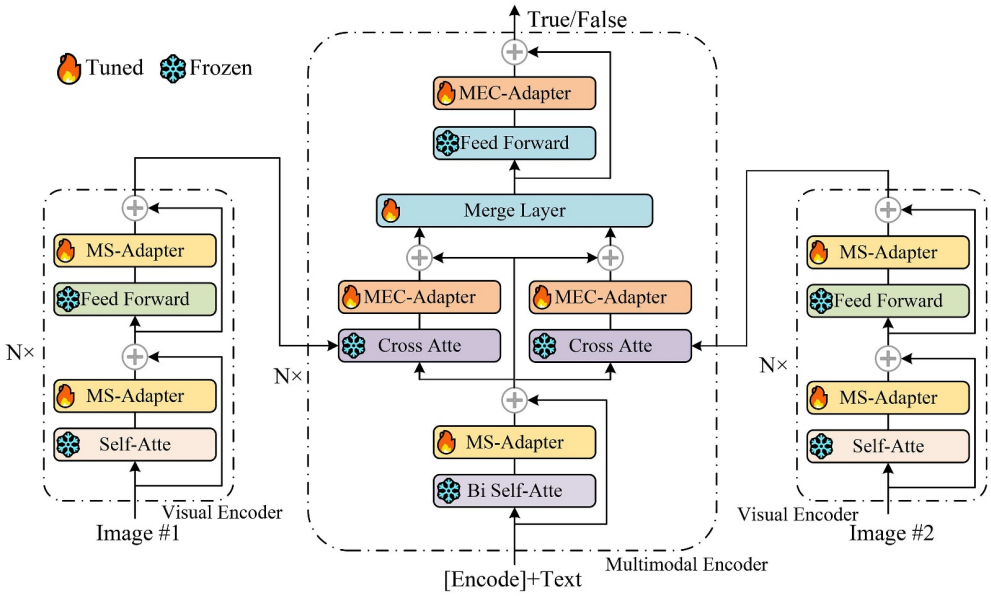

Figure 8 (Color online) Specific training framework on NLVR^2 based on the BLIP model.

Figure 9 (Color online) Visual question answering cases on the VQAv2 test set.

4.6 Visualization results and analyses

4.6.1 *Visualization results*

In this subsection, we provide several visualization results on the visual question answering task, visual reasoning task, and image captioning task, as shown in Figures 9–11. The qualitative results in Figure 9 demonstrate the effectiveness of our ModeX in visual question answering tasks. From Figure 10, our ModeX exhibits superior semantic reasoning ability and fine-grained semantic understanding ability. Specifically, in the 1st sample, the semantic information of "two women" and "at least" are captured for reasoning the right result. In the 6th sample, our ModeX also focuses on the right image by extracting the key information in the sentence, i.e., "right image". Compared with the other two methods involved parameter sharing strategy, i.e., UniAdapter [8] and ModeX w/o share, the accurate reasoning verifies the effectiveness of the modality-experts coordinated strategy. From Figure 11, we compare our method with other fine-tuning methods, including AdaptFormer [43], Adapter [2], LoRA [3], and FFT. Notably, all the models are fine-tuned with cross-entropy loss on the MSCOCO dataset and evaluated on the NoCaps validation set in a zero-shot manner, which aims to verify the generalization performance of our method. From the Figure 11, two conclusions are summarized as follows:

- Compared with the other PEFT methods, our ModeX could generate more fine-grained and diverse concepts, such as the "sliced" in the 1st instance, "bathtub" in the 2nd instance, "black and white photo" in the 4th instance.
- Compared with the FFT method, our ModeX generates nearly identical titles, such as the 1st, 2nd, and 4th instances, while enjoying fewer trainable parameters. More importantly, our ModeX achieves more accurate captions than the FFT method. Specifically, in the 6th instance, the method of FFT mistakenly recognizes the "note and pen" as "notepad", which has nearly the same letter composition

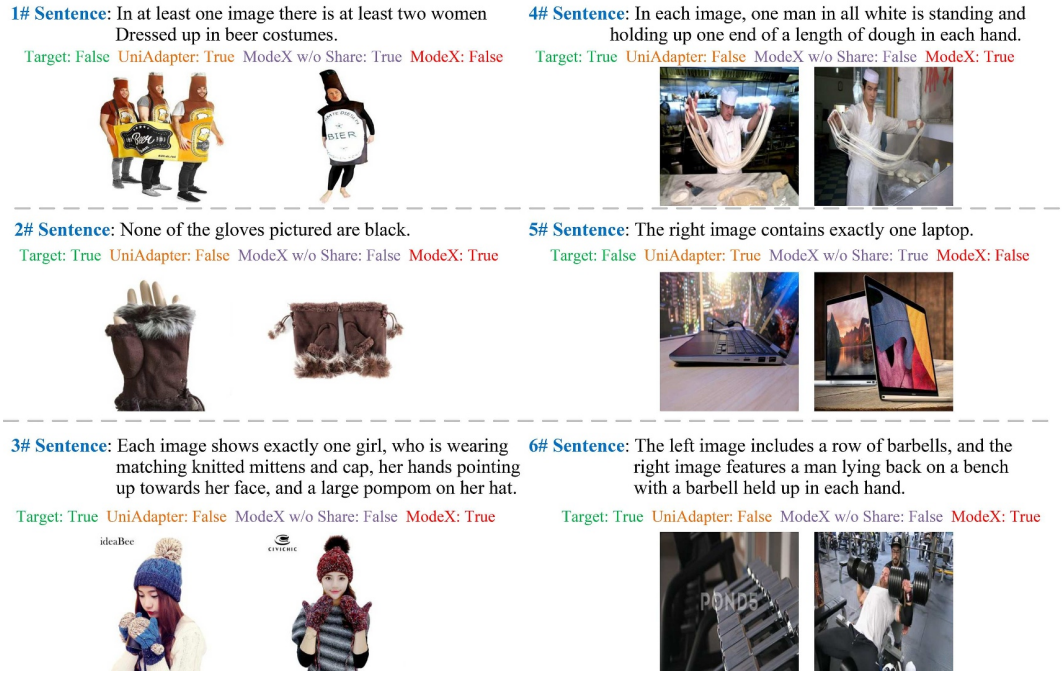

Figure 10 (Color online) Qualitative results of the reasoning results on the NLVR^2 test set, the involved comparison methods including UniAdapter [8], our ModeX w/o share, and ModeX.

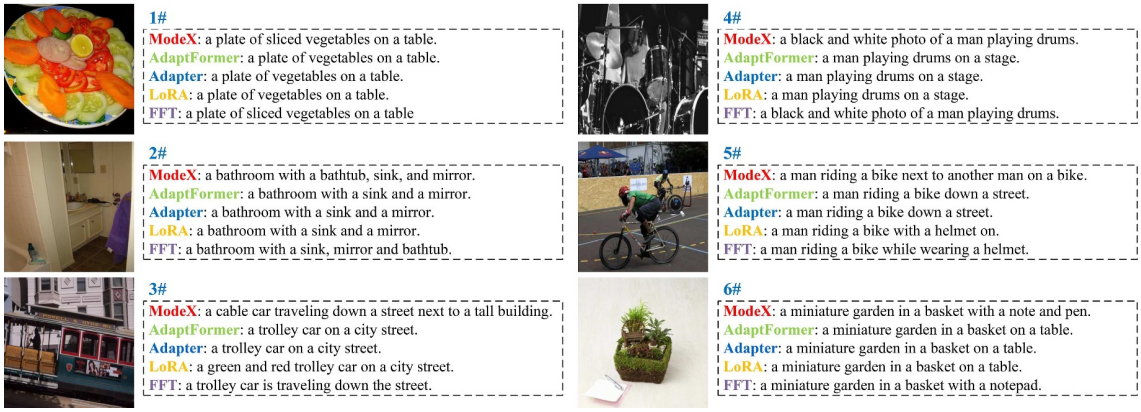

Figure 11 (Color online) Qualitative results of captioning results on the NoCaps validation set, the involved comparison methods including our ModeX, AdaptFormer [43], Adapter [2], LoRA [3], and FFT. Note that the models are exclusively fine-tuned on the MSCOCO dataset and evaluated on Nocaps with a zero-shot manner.

but vastly different semantics. By contrast, our ModeX generates the accurate caption.

4.6.2 *Negative case analyses*

In addition to the positive results, we report several negative cases and provide detailed analysis. Figure 12 illustrates the failures of two cases of image captioning and two cases of NLVR^2. The results of the first two examples generated by our ModeX are not accurate enough, and the details are not well described. For the 1# instance, our ModeX fails to capture the "white pot". For the 2# instance, "meat", "vegetables", and "wooden" are not identified by our ModeX. The above phenomena may stem from the small number of trainable parameters that hinder the visual encoder from capturing the fine-grained visual information, leading to the textual encoder or decoder's difficulty in understanding the coarse-grained visual information. As for the results of NLVR^2, i.e., 3# and 4#, we observe that our ModeX and FFT methods both predict false answers. The specific reason is that the current LMMs are difficult to deal with mathematical or counting problems, which require specific module design or training. Based on the analyses of the above cases, the entire LMMs community needs to focus on capturing fine-grained

Figure 12 (Color online) Qualitative results of negative cases on the NoCaps validation set and the NLVR^2 test set, respectively. We only provide the comparison results of FFT.

visual information and addressing specific mathematical challenges.

5 Limitations and outlooks

In this section, we discuss the limitations of our work and provide promising research directions as follows.

- As demonstrated by the analyses of the negative cases, it is evident that the LMMs community must shift its focus towards more granular capture of visual information and the development of dedicated solutions for mathematical challenges. These challenges arise since traditional LMMs often struggle to integrate the precise logical reasoning and numerical computation. By addressing these specific requirements, future advancements in LMMs will achieve greater accuracy and reliability in a broader range of applications.
- Although the advanced LMMs have achieved significant improvements in OCR-related tasks with stronger vision encoder or specific algorithms, such as the dynamic high resolution [70] and the multimodal rotary position embedding [71], they still face limitations due to large parameter sizes and prohibitive computational demands, limiting real-time applications on edge devices. In the field of PEFT, we prefer to choose offline specialist models as reward models to efficiently fine-tune the LMMs and inject domain-specific knowledge. Additionally, efficiently fine-tuning the LMMs with additional high-quality bilingual datasets will contribute to recognizing more fine-grained text content, especially the ChartQA [68] and the DocVQA [69].
- We focus on substantially reducing the number of trainable parameters and per-task storage for alleviating the requirements of fine-tuning the current large foundation models. However, like most of the PEFT methods in the realm of NLP or CV, the combination of our proposed method with VLP models still requires quite a few computing resources and time consumption during training, despite significantly reduces the memory requirement than that of FFT. Thus, more efficient techniques should be designed to further address this issue, which enables more researchers leverage and even fine-tune the ever-increasing large foundation models under constrained budget.

6 Conclusion

This paper has proposed a ModeX method for effectively adapting various LMMs to downstream tasks, which alleviates the limitations of the existing PEFT methods, specifically in terms of low performance and poor compatibility. ModeX showcases an effective, plug-and-play, and lightweight adaptation method tailored to various LMMs. We empirically divide the components of LMMs into UM-MSA/FFN and CM-MSA/FFN, which aims at reformulating a unified classification for management. Particularly, we achieve ModeX by dedicatedly coordinating the modality-specific and modality-mixture experts within a novel routing algorithm. Extensive experiments demonstrate that ModeX exhibits comparable or superior performance to both FFT and state-of-the-art PEFT methods, while demonstrating high compatibility with diverse LMMs. Moreover, our ModeX also showcases superior training efficiency in both the number of training parameters and training duration, further underlining its contribution to the community.

In summary, this study provides a robust and versatile parameter adaptation method for the evolving landscape of LMMs. The successful application of ModeX suggests its potential as a pivotal direction in future LMM research, offering potent support for a broader range of application scenarios.

Acknowledgements This work was supported by National Key Research and Development Program of China (Grant No. 2022ZD0160403) and National Natural Science Foundation of China (Grant No. 62176178).

References

1 Li X L, Liang P. Prefix-tuning: optimizing continuous prompts for generation. In: Proceedings of Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing, Bangkok, 2021. 4582–4597
2 Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter-efficient transfer learning for NLP. In: Proceedings of International Conference on Machine Learning, Los Angeles, 2019. 2790–2799
3 Hu E J, Shen Y L, Wallis P, et al. LoRA: low-rank adaptation of large language models. In: Proceedings of International Conference on Learning Representations, 2022
4 Jia M L, Tang L M, Chen B C, et al. Visual prompt tuning. In: Proceedings of European Conference on Computer Vision, Tel Aviv, 2022. 709–727
5 Yang T, Zhu Y, Xie Y S, et al. AIM: adapting image models for efficient video understanding. In: Proceedings of International Conference on Learning Representations, Kigali, 2023
6 Jiang H J, Zhang J K, Huang R, et al. Cross-Modal adapter for text-video retrieval. 2022. ArXiv:2211.09623
7 Long Z, Killick G, McCreadie R, et al. MultiWay-Adapter: adapting multimodal large language models for scalable image-text retrieval. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Seoul, 2024. 6580–6584
8 Lu H, Huo Y, Yang G, et al. UniAdapter: unified parameter-efficient transfer learning for cross-modal modeling. In: Proceedings of International Conference on Learning Representations, Vienna, 2024
9 Wang H X, Yang X L, Chang J L, et al. Parameter-efficient tuning of large-scale multimodal foundation model. In: Proceedings of Advances in Neural Information Processing Systems, New Orleans, 2023. 15752–15774
10 Yuan Y, Zhan Y, Xiong Z. Parameter-efficient transfer learning for remote sensing image-text retrieval. IEEE Trans Geosci Remote Sens, 2023, 61: 1–14
11 Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: Proceedings of International Conference on Learning Representations, 2021
12 Kenton J D M W C, Toutanova L K. BERT: pretraining of deep bidirectional transformers for language understanding. In: Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, 2019. 4171–4186
13 Li J N, Li D X, Xiong C M, et al. BLIP: bootstrapping language-image pretraining for unified vision-language understanding and generation. In: Proceedings of International Conference on Machine Learning, Baltimore, 2022. 12888–12900
14 Li J N, Li D X, Savarese S, et al. BLIP-2: bootstrapping language-image pretraining with frozen image encoders and large language models. In: Proceedings of International Conference on Machine Learning, Honolulu, 2023. 19730–19742
15 Zhang S S, Roller S, Goyal N, et al. OPT: open pre-trained transformer language models. 2022. ArXiv:2205.01068
16 Wang W, Bao H, Dong L, et al. Image as a foreign language: BEIT pretraining for vision and vision-language tasks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Vancouver, 2023. 19175–19186
17 Changpinyo S, Sharma P, Ding N, et al. Conceptual 12M: pushing web-scale image-text pretraining to recognize long-tail visual concepts. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2021. 3558–3568
18 Schuhmann C, Vencu R, Beaumont R, et al. LAION-400M: open dataset of CLIP-filtered 400 million image-text pairs. 2021. ArXiv:2111.02114
19 Ji Z, Chen K X, He Y Q, et al. Heterogeneous memory enhanced graph reasoning network for cross-modal retrieval. Sci China Inf Sci, 2022, 65: 172104
20 Zhang Y, Ji Z, Pang Y W, et al. Consensus knowledge exploitation for partial query based image retrieval. IEEE Trans Circ Syst Video Technol, 2023, 33: 7900–7913
21 Zhang Y, Ji Z, Wang D, et al. USER: unified semantic enhancement with momentum contrast for image-text retrieval. IEEE Trans Image Process, 2024, 33: 595–609
22 Ji Z, Meng C, Zhang Y, et al. Knowledge-aided momentum contrastive learning for remote-sensing image text retrieval. IEEE Trans Geosci Remote Sens, 2023, 61: 1–13
23 Ji Z, Li Z, Zhang Y, et al. Hierarchical matching and reasoning for multi-query image retrieval. Neural Netws, 2024, 173: 106200
24 Wang Z R, Yu J H, Yu A W, et al. SimVLM: simple visual language model pretraining with weak supervision. In: Proceedings of International Conference on Learning Representations, 2021
25 Yang Y, Bao R, Guo W L, et al. Deep visual-linguistic fusion network considering cross-modal inconsistency for rumor detection. Sci China Inf Sci, 2023, 66: 222102
26 Antol S, Agrawal A, Lu J, et al. VQA: visual question answering. In: Proceedings of IEEE International Conference on Computer Vision, Santiago, 2015. 2425–2433
27 Suhr A, Zhou S, Zhang A, et al. A corpus for reasoning about natural language grounded in photographs. In: Proceedings of Annual Meeting of the Association for Computational Linguistics, Bangkok, 2019. 6418–6428
28 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of International Conference on Machine Learning, 2021. 8748–8763
29 Jia C, Yang Y F, Xia Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision. In: Proceedings of International Conference on Machine Learning, 2021. 4904–4916
30 Chen Y C, Li L J, Yu L C, et al. UNITER: universal image-text representation learning. In: Proceedings of European Conference on Computer Vision, Glasgow, 2020. 104–120
31 Li X J, Yin X, Li C Y, et al. Oscar: object-semantics aligned pretraining for vision-language tasks. In: Proceedings of European Conference on Computer Vision, Glasgow, 2020. 121–137
32 Kim W, Son B, Kim I. ViLT: vision-and-language transformer without convolution or region supervision. In: Proceedings of International Conference on Machine Learning, 2021. 5583–5594
33 Li J N, Selvaraju R, Gotmare A, et al. Align before fuse: vision and language representation learning with momentum distillation. In: Proceedings of Advances in Neural Information Processing Systems, 2021. 9694–9705
34 Zhou K, Yang J, Loy C C, et al. Learning to prompt for vision-language models. Int J Comput Vis, 2022, 130: 2337–2348

35 Zhou K, Yang J, Loy C C, et al. Conditional prompt learning for vision-language models. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, 2022. 16816–16825
36 Khattak M U, Rasheed H, Maaz M, et al. MaPLe: multimodal prompt learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Vancouver, 2023. 19113–19122
37 Yang H, Lin J Y, Yang A, et al. Prompt tuning for generative multimodal pretrained models. 2022. ArXiv:2208.02532
38 Hu Z Y, Li Y Y, Lyu M R, et al. VL-PET: vision-and-language parameter-efficient tuning via granularity control. In: Proceedings of IEEE International Conference on Computer Vision, Paris, 2023. 3010–3020
39 Sung Y L, Cho J, Bansal M. Vl-Adapter: parameter-efficient transfer learning for vision-and-language tasks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, 2022. 5227–5237
40 Liu H T, Li C Y, Wu Q Y, et al. Visual instruction tuning. In: Proceedings of Advances in Neural Information Processing Systems, New Orleans, 2023. 34892–34916
41 Cho J, Lei J, Tan H, et al. Unifying vision-and-language tasks via text generation. In: Proceedings of International Conference on Machine Learning, 2021. 1931–1942
42 Luo G, Zhou Y Y, Ren T H, et al. Cheap and quick: efficient vision-language instruction tuning for large language models. In: Proceedings of Advances in Neural Information Processing Systems, New Orleans, 2023. 29615–29627
43 Chen S F, Ge C J, Tong Z, et al. AdaptFormer: adapting vision transformers for scalable visual recognition. In: Proceedings of Advances in Neural Information Processing Systems, New Orleans, 2022. 16664–16678
44 Chen X L, Fang H, Lin T Y, et al. Microsoft COCO captions: data collection and evaluation server. 2015. ArXiv:1504.00325
45 Plummer B A, Wang L W, Cervantes C M, et al. Flickr30K entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of IEEE International Conference on Computer Vision, Santiago, 2015. 2641–2649
46 Xie N, Lai F, Doran D, et al. Visual entailment: a novel task for fine-grained image understanding. 2019. ArXiv:1901.06706
47 Goyal Y, Khot T, Summers-Stay D, et al. Making the V in VQA matter: elevating the role of image understanding in visual question answering. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, 2017. 6904–6913
48 Andrej K, Li F F. Deep visual-semantic alignments for generating image descriptions. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 3128–3137
49 Agrawal H, Desai K, Wang Y F, et al. NoCaps: novel object captioning at scale. In: Proceedings of IEEE International Conference on Computer Vision, Seoul, 2019. 8948–8957
50 Krishna R, Zhu Y, Groth O, et al. Visual Genome: connecting language and vision using crowdsourced dense image annotations. Int J Comput Vis, 2017, 123: 32–73
51 Hudson D A, Manning C D. GQA: a new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 6700–6709
52 Lu P, Mishra S, Xia T, et al. Learn to explain: multimodal reasoning via thought chains for science question answering. In: Proceedings of Advances in Neural Information Processing Systems, New Orleans, 2022. 2507–2521
53 Fu C Y, Chen P X, Shen Y H, et al. MME: a comprehensive evaluation benchmark for multimodal large language models. 2023. ArXiv:2306.13394
54 Yu W, Yang Z, Li L, et al. MM-Vet: evaluating large multimodal models for integrated capabilities. In: Proceedings of International Conference on Machine Learning, Vienna, 2024
55 Li Y, Du Y, Zhou K, et al. Evaluating object hallucination in large vision-language models. In: Proceedings of Empirical Methods in Natural Language Processing, Sentosa, 2023. 292–305
56 Gurari D, Li Q, Stangl A J, et al. Vizwiz grand challenge: answering visual questions from blind people. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 3608–3617
57 Liu Y, Duan H, Zhang Y, et al. MMBench: is your multi-modal model an all-around player? 2023. ArXiv:2307.06281
58 Singh A, Natarajan V, Shah M, et al. Towards VQA models that can read. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 8317–8326
59 Liu H T, Li C Y, Li Y H, et al. Improved baselines with visual instruction tuning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Seattle, 2024. 26296–26306
60 Li D X, Li J N, Le H, et al. LAVIS: a one-stop library for language-vision intelligence. In: Proceedings of Annual Meeting of the Association for Computational Linguistics, Toronto, 2023. 31–41
61 He K M, Zhang X Y, Ren S Q, et al. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of IEEE International Conference on Computer Vision, Santiago, 2015. 1026–1034
62 Gan Z, Chen Y C, Li L J, et al. Large-scale adversarial training for vision-and-language representation learning. In: Proceedings of Advances in Neural Information Processing Systems, New Orleans, 2020. 6616–6628
63 Liu X, Ji K X, Fu Y C, et al. P-Tuning v2: prompt tuning can be comparable to fine-tuning universally across scales and tasks. In: Proceedings of Annual Meeting of the Association for Computational Linguistics, Dublin, 2022. 61–68
64 Zhang R R, Han J M, Liu C, et al. LLaMA-Adapter: efficient fine-tuning of language models with zero-init attention. In: Proceedings of International Conference on Learning Representations, Vienna, 2024
65 Zhang P C, Li X J, Hu X W, et al. VinVL: revisiting visual representations in vision-language models. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2021. 5579–5588
66 Tan H, Bansal M. LXMERT: learning cross-modality encoder representations from transformers. In: Proceedings of Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing, 2019. 5100–5111
67 Chiang W L, Li Z H, Lin Z, et al. Vicuna: an open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. 2023. https://lmsys.org/blog/2023-03-30-vicuna/
68 Masry A, Do X L, Tan J Q, et al. ChartQA: a benchmark for question answering about charts with visual and logical reasoning. In: Proceedings of Findings of the Association for Computational Linguistics, 2022. 2263–2279
69 Mathew M, Karatzas D, Jawahar C V. DocVQA: a dataset for VQA on document images. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision, 2021. 2200–2209
70 Chen Z, Wang W Y, Tian H, et al. How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites. 2024. ArXiv:2404.16821
71 Wang P, Bai S, Tan S N, et al. Qwen2-VL: enhancing vision-language model's perception of the world at any resolution. 2024. ArXiv:2409.12191