PILL: Plug into LLM with Adapter Expert and Attention Gate

Yuyu Yin[a,*], Fangyuan Zhang[a,*], Zhengyuan Wu[a], Qibo Qiu[b], Tingting Liang[a,**], Xin Zhang[a]

[a]*College of Computer Science & Technology, Hangzhou Dianzi University, Hangzhou*
[b]*ZhejiangLab*

Abstract

Due to the remarkable capabilities of powerful Large Language Models (LLMs) in effectively following instructions, significant progress recently has been made in the development of Vision Language Models (VLMs), expanding the capabilities of LLMs in multi-modal learning and enabling them to be fine-tuned in a parameter-efficient manner. Promising as the existing pre-trained vision-language models might be, there are two intractable issues including modal entanglement and insufficient capability of LLMs in handling visual information. To tackle these two issues, we introduce an architecture called **PILL** which **P**lugs **I**nto **LL**M with adapter expert and attention gate. To maintain the capability of the original LLMs and the more progressive injection of visual modality, one Modality-Attention-Gating (MAG) module is introduced, enabling adaptive control of the contribution of modality tokens to the overall representation. Specifically, one Mixture-of-Modality-Adapter-Expert (MoMAE) module is proposed to independently handle different modalities with the dedicated adapters. In addition, further improvement is made to the adapter to enhance its learning and expressive capabilities. We adopt a two-stage training paradigm to optimize different modules of our model. Experimental results demonstrate that our approach exhibits competitive performance compared to other mainstream methods for modality fusion with much lower resources. We provide free access to the code and models at https://github.com/DsaltYfish/PILL.

*Indicates equal contribution
**Corresponding author
Email address: liangtt@hdu.edu.cn (Tingting Liang)

Keywords: Modality Entanglement, Adapter Expert, Modality Attention Gate, Multi-Modal

1. Introduction

The long-term goal of artificial intelligence is to utilize knowledge in a human-like manner for tasks such as reasoning, thinking, analysis, and decision-making. With the remarkable instruction-following ability and astonishing comprehension of human language exhibited by large language models (LLMs) [1, 2], Universal Visual Language Models (VLMs) [3, 4] have made significant progress in the field of AI's multimodal capabilities. In order to learn the more representative multi-modal features, large-scale pre-trained vision-language models have gained tremendous attention on a variety of cross-modal tasks [5, 6], whose objective is acquiring the cross-modal representations by aligning visual and linguistic modalities with a large quantity of image-text pairs.

Despite having modality information aligned with language, fully fine-tuning such large-scale pre-trained VLMs requires high time and resource costs. Recently, one straightforward approach is to utilize the Parameter-Efficient Fine-Tuning (PEFT) method to enable the LLMs to learn how to handle the multi-modal information [7, 8]. Promising as the previous PEFT-based VLMs might be, the internal structures of the underlying large language models that share the same single network for multi-modality fusion suffer from the modal entanglement issue, where different modalities interfere with each other, leading to the dominance of one modality and the suppression of others. This issue poses a significant challenge, hindering the effective integration and utilization of multimodal information. In addition, the single networks in the most previous work generally fuse the multi-modal information with self-attention or cross-attention mechanism which handles each modality in the same manner. However, the LLMs lack the capability of directly handling the visual information, which might potentially lead to injecting inaccurate visual signals into the textual representations. To address this limitation, a more principled approach is required to enable the harmonious fusion of multiple modalities while preserving their respective strengths and characteristics.

To alleviate the aforementioned issues, this work proposes a new multimodal frame-

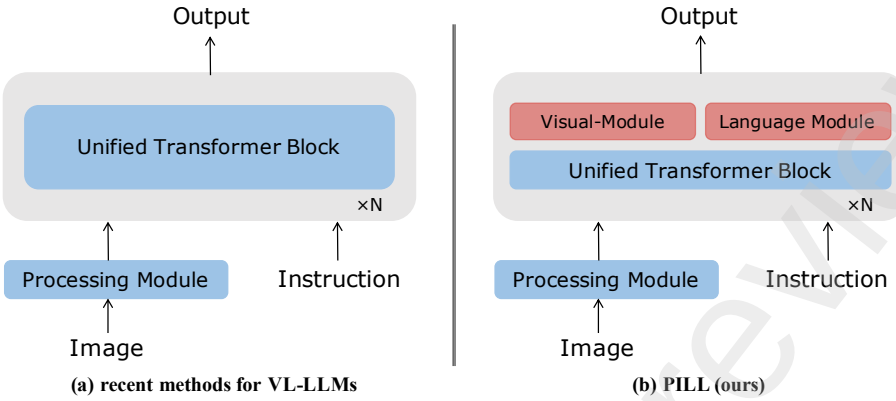

Figure 1: Comparison between recent VLMs methods and our approach. (a) represents the recent VL-LLMs method. The Processing Module represents the modules used for image processing, and transfer image information to the LLMs. (b) represents our approach. We keep the image processing modules consistent with other methods and introduce a dedicated module within the LLMs to handle image information.

work that Plugs Into LLM with adapter expert and attention gate (PILL), aiming to bridging the gap between pre-trained language models and effective multimodal understanding. Different from the previous works that use the shared network for multi-modality fusion (Figure 1(a)), we allocate dedicated modules as shown in Figure 1(b), which respectively handle the visual and textual modalities. Specifically, aiming at the insufficient capability of LLMs in handling visual information, we introduce one Modality-Attention-Gating (MAG) module which implements the gradual infusion of visual signals into textual modality with the learned attention weights, guaranteeing the ability of language model by keeping the consistency between the outputs and those of the original language model during initial training. In this manner, the MAG can stabilize the training process and prevent visual information from interfering with the LLM's text modeling phase. For the challenge of modal entanglement, a Mixture-of-Modality-Adapter-Expert (MoMAE) is proposed, which equips each modality with dedicated FFNs to handle different modality inputs. Furthermore, we refine the classical adapter by introducing the SwiGLU-Adapter [9] to lead the improved performance.

To ensure stable training of our model, we adopt a two-stage training strategy. In

the first stage, we train the projection layer and V-Adapter separately to align visual and textual information, and enable the LLM to learn to process visual knowledge. In the second stage, we unlock all the proposed modules for downstream fine-tuning. During both training stages, the LLM and Visual Prompt Generation (VPG) [10] modules remain frozen. A notable advantage of our training method is its efficiency since we only train the inserted modules. Our model can be efficiently trained on a single A6000 GPU, allowing for rapid training completion. The experimental results demonstrate that our model is able to achieve competitive performance compared to other methods with much lower resources.

In summary, our contributions can be summarized as follows:

1. We propose a parameter-efficient vision-language modality learning framework which plugs into LLM with adapter expert and attention gate (PILL). It consists of one MoMAE module for alleviating the modal entanglement issue and one MAG module for better modal fusion with the dynamic adjustment of visual signal injection into the LLM.

2. PILL demonstrates superior training and computational efficiency, as our experiments can be completed on a single A6000 GPU.

3. Experimental results demonstrate that PILL exhibits superior efficiency and competitive performance compared to existing multi-modal LLMs. It also shows the significant potential for handling multi-modal instructions while preserve the dialog capabilities.

The remainder of this paper is structured as follows. Section 2 introduces some previous works about large-scale multimodal models and Multi-Modal Instruction-Tuning in PEFT. Section 3 gives a specific statement for the problem investigated in this work and proposed MoMAE and MAG are described. Section 4 reports the experimental results. Section 5 discusses the Limitation and Future work. Section 6 draws the conclusion.

2. Related Work

2.1. Multimodal Fusion

By aligning and integrating different modalities, researchers aim to unlock the full potential of multimodal representations. Early efforts in this direction such as VL-BERT [11], ViL-BERT [12], MCAN [13], LXMERT [14], ERNIE-ViL [15, 16], and Rosita [17] extract key information from images using object detectors and process it together with text representation in transformer blocks [18]. Furthermore, CLVIN [19] refines the encoder-decoder architecture based on the MCAN framework to jointly encode visual and textual representations, KVL-BERT [20]incorporates commonsense knowledge into the cross-modal BERT model to enhance the model's understanding of entity relationships and semantic information. [16] introduces a gated attention mechanism to further refine and optimize visual information, and IASSM [21] combines inferential attention and semantic space mapping. These approaches leverage the self-attention mechanism to dynamically attend to and aggregate relevant information from both modalities, enabling more seamless integration and facilitating bidirectional cross-modal interactions. Despite their promising performance, these methods rely on explicit object detection and region proposals, which may limit their ability to capture more holistic and contextual visual information and introduce additional computational overhead.

2.2. Large-Scale Multimodal Models

Large-scale pretraining has played a crucial role in the field of multimodal learning, enabling more robust and versatile AI systems capable of tackling complex real-world tasks that involve both visual and linguistic components. Following the impressive success of CLIP [22] in image-text retrieval, contrastive learning has gained attention in the multimodal domain. Methods like ALBEF [23], BLIP [24], VLMO [25], BEiT-3 [26], and CoCa [27] utilize contrastive learning to align image and text features, demonstrating the benefits of such alignment. Subsequently, with advancements in large-scale model techniques, methods like ClipCap [28], PaLi [29], PaLM [4], KOSMOS [30, 31, 32], BLIP-2 [10] incorporating features from images specifically processed by

ViT [33] as prompts in the LM's input. More recently, with the rise of LLMs, researchers have focused on leveraging the powerful capabilities of LLMs and combining them with visual information. Mini-GPT4 [34] and LLaVA [35, 36] have discovered that a projection layer can effectively project visual information into the textual space of LLMs, and they train only this projection layer using large-scale pretraining data. This allows for efficient integration of visual modalities into the well-established linguistic representations of LLMs.

2.3. Multi-Modal Instruction-Tuning in PEFT

To address the high cost associated with full fine-tuning, the Parameter-Efficient Fine-Tuning (PEFT) technique has emerged as an alternative. By fine-tuning only a small portion of the model's parameters, PEFT techniques can significantly reduce the computational cost of fine-tuning while maintaining the model's generalization capability. This makes it possible to rapidly adapt large pretrained models to downstream tasks in resource-constrained environments. We highlight three commonly used PEFT techniques and discuss their applications in various methods. Firstly, the Adapter [37] approach has been explored in VL-Adapter [38]. Extensive experiments have demonstrated the advantages of adapters over other PEFT methods in the multimodal setting. MAGMA [39] introduces adapters within the LM based on a frozen model. LAVIN [8] adopts a similar approach by treating single-modal and multimodal inputs as separate tasks and utilizing MMA (Mixture-of-Modality Adapter) for multi-modality task. The key distinction between our MoMAE approach and LAVIN is that we focus on modality tokens, while LAVIN focuses on tasks. Next, the application of LoRA [40] in the multimodal context is noteworthy. MultiModal-GPT [41] adopts a similar architecture to Flamingo [42] and incorporates LoRA for LLM fine-tuning. Visual-ChatGLM [43], mPLUG-Owl [44], and LAMM [45] also employ LoRA for LM fine-tuning after pretraining with VPG (Visual Prompt Generation). Finally, the prefix-tuning techniques [46], exemplified by LLAMA-Adapter [7, 47], involve adding the image representation from the visual encoder to the prefix tokens and processing it along with the text tokens in the LM layers.

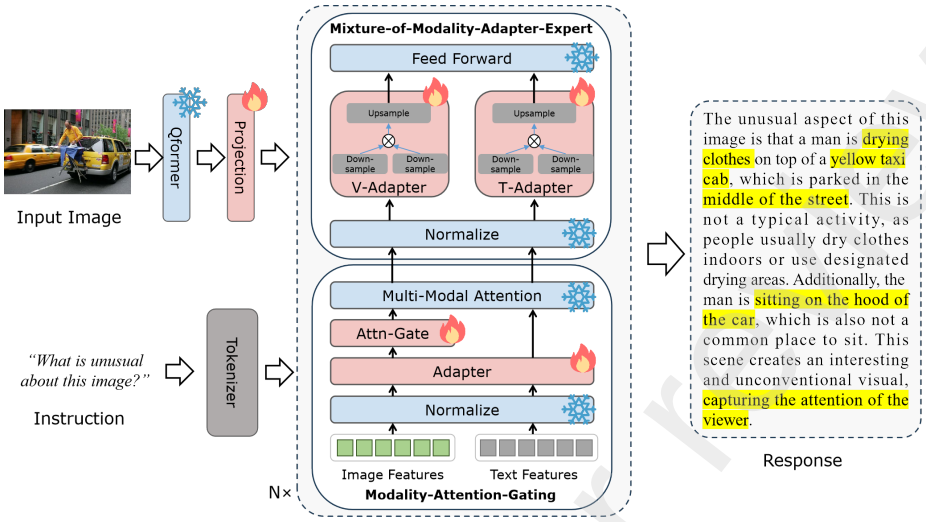

Figure 2: The overview of the architecture of PILL and two module of the PILL: Mixture-of-Modality-Adapter-Expert (MoMAE) and Modality-Attention-Gating (MAG). In PILL, the MoMAE are employed to handle tokens from different modalities. During fine-tuning phase, MAG is used for coordinate the weights of other modalities

3. Method

In this section, we firstly present the overall model architecture of PILL (Sec.3.1). Then we introduce the Mixture-of-Modality-Adapter-Expert (MoMAE) (Sec.3.3), a dedicated module designed to handle different modalities of information effectively. Furthermore, we propose the Modality-Attention-Gating (MAG) mechanism (Sec.3.2), which enables the LLM to progressively integrate visual information during fine-tuning. Additionally, we describe the SwiGLU-Adapter (Sec.3.4), an improved adapter design that enhances adapter learning and expressive capability. After a detailed introduction of these modules, we will proceed to discuss the training process and objectives of the PILL model (Sec. 3.5). An overview of the PILL framework is depicted in Figure 2.

3.1. Model Architecture

Given a sample containing a set of images and texts, we first process each image in the image set using the Q-former module from BLIP-2 [10] to extract image fea-

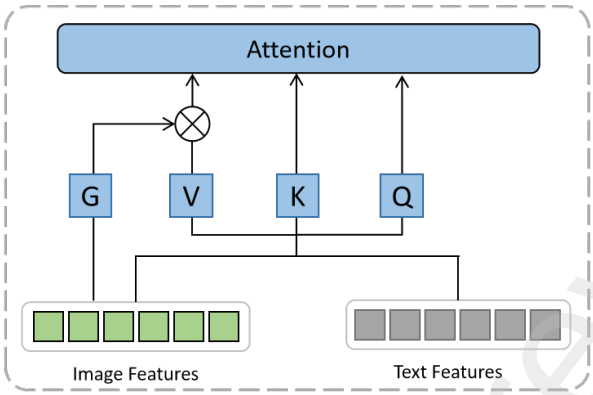

Figure 3: Illustration of the Modality-Attention-Gating (MAG)

tures. Q-former takes a fixed number of K learnable queries to interact with image features. We then project the dimension of these features to match the dimension of LLaMA using a projection layer, and obtain $V_j = \{v_{j1}, ..., v_{jK}\}$ which represents the visual prompts for the j-th interleaved image. For the texts, using the tokenizer within LLaMA, tokens are encoded into embeddings as h_i, which represents the i-th textual token. We denote the visual features of M images and N token embeddings by $H^v = \{v_{11}, ..., v_{1K}, ..., v_{M1}, ..., v_{MK}\}$ and $H^t = \{h_1, h_2, ..., h_N\}$. Next, these encoded features are concatenated based on the positions of the original texts and images in the sample and fed into PILL. The input features can be represented as $H = \{h_1, h_2, ..., v_{11}, ..., v_{1K}, ..., h_i, ..., v_{j1}, ..., v_{jK}, ..., h_N\}$. The concatenation of visual and textual features enables PILL to process multimodal inputs in a unified manner, leveraging the powerful language understanding capabilities of LLaMA while seamlessly integrating visual information.

3.2. Modality-Attention-Gating

Due to large language models (LLMs) being pretrained solely on linguistic corpora without access to visual data, naively concatenating raw image features to the text input may disrupt the well-calibrated parameters tuned for language modeling tasks. This direct fusion approach can potentially introduce interference and hinder the model's ability to effectively leverage its pre-trained linguistic knowledge.

Furthermore, as mentioned earlier, the modal entanglement issue poses a significant

challenge, where different modalities can interfere with each other, leading to the dominance of one modality and the suppression of others. To address this issue and facilitate effective cross-modal integration, we aim to introduce a mechanism that allows for progressive and controlled fusion of visual information during the fine-tuning stage.

To this end, we propose the Modality-Attention-Gating (MAG) mechanism. Specifically, we apply MAG to adjust the attention weights associated with other modality information, aiming to allow each modal information to play its respective role in different transformer layers. By gradually exposing the language model to visual signals during fine-tuning, MAG mitigates the modal entanglement issue and enables the model to effectively adapt to and integrate visual information without compromising its pre-trained linguistic capabilities.

As is shown in Figure 3, we specify $G : \mathbb{R}^{d_{dim}} \rightarrow \mathbb{R}^{n_{heads}}$. Given vision modality tokens $V(H_{in}^{v})$ and text modality tokens $V(H_{in}^{t})$, our MAG module is computed as follows:

$$\begin{aligned} G_v(H) &= Tanh(G(H_{in}^{v}))V(H_{in}^{v}) \\ G_t(H) &= V(H_{in}^{t}) \\ S(H) &= Softmax(\frac{Q(H_{in})K(H_{in})}{\sqrt{d_{head}}}) \\ MAG(H) &= [G_v(H), G_t(H)] \odot S(H) \end{aligned} \tag{1}$$

In attention layer, H is reshaped into (n_{heads}, T, d_{head}). The role of $G_v(H)$ is to multiply the attention heads of visual modality by a gating weight $G(H_{in}^{v})$, which is learned by a fully-connected layer and initialized to 0. We multiply the gated weights to the $V(H_{in}^{v})$ in self-attention, which is equivalent to reweighting the attention scores corresponding to the visual modality within the attention matrix. It mitigates the influence introduced by partially trained visual information and enables progressive fusion. Note that for every visual token within the same layer, they undergo the same gate. By providing a controlled and progressive fusion mechanism, MAG facilitates a smooth transition for the model, allowing it to leverage its linguistic knowledge while effectively assimilating visual information. This approach not only preserves the model's pre-trained capabilities but also enhances its multimodal understanding and

reasoning abilities through effective cross-modal integration.

3.3. Mixture-of-Modality-Adapter-Expert

Due to the substantial discrepancies across modalities, there exists cross-modal interference, where the dominant modality governs the output and suppresses the training of weaker modalities. Single-stream base models struggle to strike a balance between the benefits of cross-modal collaboration and cross-modal influence. To address this modal entanglement issue [6, 48], we propose a universal multimodal module, named Mixture-of-Modality-Adapter-Expert (MoMAE).

MoMAE introduces a Mixture-of-Modality-Adapter-Expert as an alternative solution to the modality expert feed-forward networks for encoding different modalities. As shown in Figure 2, MoMAE is capable of capturing and processing modality-specific information by switching to different modality adapter experts. We employ a vision adapter expert for encoding images H^v, and a language adapter expert for encoding text H^t, which can be formulated as:

$$\begin{aligned} H^v_{out} &= Adapter_v(H^v_{in}) \\ H^t_{out} &= Adapter_t(H^t_{in}), \end{aligned} \tag{2}$$

where H^v and H^t represent the hidden states of visual and textual modalities, respectively. $Adapter_v$ and $Adapter_t$ contain separated, modality-focused parameters. They share the same lightweight SwiGLU-Adapter architecture, which will be introduced in Sec.3.4

We employ MoMAE to enable LLM to learn knowledge for handling visual features. Additionally, in our experiments, we observed that the variance of visual tokens is typically significantly larger than that of textual tokens. Employing a dedicated module to handle visual tokens may be a prudent choice. Furthermore, MoMAE alleviates the modal entanglement issue by maintaining separate parameter spaces for different modalities, preventing the dominance of a single modality during training. By leveraging modality-specific adapter experts, MoMAE enables the model to flexibly integrate and process information from various modalities, enhancing its multimodal capabilities. Additionally, due to the adoption of the Adapter strategy for learning, the weights of the

original LLaMA are retained, allowing the alignment of visual and textual modalities in the same representation space.

3.4. SwiGLU-Adapter

The Gated Linear Unit (GLU) [9] structure has gained wide application in large language models (LLMs) such as LLaMA [49, 2] and PaLM [4]. Empirical evidence substantiates that, compared to conventional activation functions such as ReLU and GeLU, the GLU variant yields appreciable performance gains.

The classical Adapter architecture consists of a downsampling operation, activation function, upsampling, and residual connection. However, the classical Adapter design may not fully leverage the expressive power of the GLU structure. To synthesize the benefits of Adapters and GLU, while enhancing the learning capacity and representational ability of the model, we propose the SwiGLU-Adapter. The SwiGLU-Adapter aims to introduce a more flexible and effective gating mechanism, enabling the model to better capture and integrate multimodal information.

We modify the structure of the Adapter by incorporating two downsampling operations, where one of the downsampling outputs is passed through an activation function. The outputs are then multiplied element-wise and followed by an upsampling operation. We represent this transformation using the following equations:

$$\begin{aligned} H_{down1} &= SiLU(D_1(H_{in})) \\ H_{down2} &= D_2(H_{in}) \\ H_{down} &= H_{down1} \otimes H_{down2} \\ H_{up} &= U(H_{down}) \\ H_{out} &= H_{up} \oplus H_{in} \end{aligned} \tag{3}$$

By introducing an additional gating mechanism and nonlinear transformation, the SwiGLU-Adapter offers improved expressive capability and adaptability, allowing the model to better capture intricate multimodal interactions. This design choice is motivated by the need to enhance the learning capacity of the model, enabling it to effectively integrate and process information from various modalities, ultimately leading to improved multimodal representations and performance.

3.5. Stagewise Training

Since the gradients for the visual modality are propagated solely through the attention layers, directly applying PILL to downstream tasks may result in insufficient training of the vision experts. Consequently, we propose a two-stage training approach to strike a balance between leveraging the pre-trained multimodal representations and adapting the model to the unique characteristics of each downstream task. We ensure that the model acquires a strong multimodal foundation while retaining the flexibility to specialize for specific applications, ultimately enhancing its performance and generalization capabilities across a wide range of multimodal tasks.

Our training process consists of the classic pre-training and downstream task fine-tuning stages, where different modules are trained at each stage. For training objective, both stage train on loss which is an auto-regressive manner. Given input X_{text}, X_{image} and X_{answer}, we express our training objective using the following equation:

$$p_t = \prod_{s=1}^{S} p(X_s | X_{text}, X_{image}, X_{answer,<s}; \theta), \tag{4}$$

where p_t, S, and θ denote the probabilities of the predicted word, the length of answer sequence, the trainable parameters, respectively. We maximize the likelihood of p_t with trainable parameters θ in both training stage. The following sections will provide a detailed description of each stage.

Stage 1: Pretraining for Modality Alignment.

The pre-training stage plays a crucial role in endowing the model with a strong understanding of the multimodal data distribution and acquiring effective cross-modal representations. By training on a large-scale corpus of image-text pairs, the model learns to align and integrate information from different modalities, laying the foundation for robust multimodal understanding and reasoning. This stage is essential for developing a versatile and generalizable model capable of handling diverse multimodal tasks.

In this stage, we set $\theta = \{A_v, P\}$, where A_v is visual adapter and P is the projection layers. By doing so, the visual features extracted from the Q-former can align with the pre-trained LLM word embeddings, allowing the FFN layers in LLM to learn preliminary knowledge in processing visual tokens.

It is noteworthy that although the MAG module is tailored exclusively for the visual modality, it is not trained during the pretraining stage. This design ensures that the visual expert and projection layers receive sufficient training, as the inclusion of MAG during pretraining would disrupt the gradient flow within the visual modality.

Stage 2: Downstream Task Fine-tuning.

The fine-tuning stage allows the model to refine its cross-modal fusion capabilities, enabling it to effectively leverage visual cues and linguistic context in a task-specific manner. We train the model to respond to instructions and images like a real AI assistant. Therefore, we train all the parameters of the Adapters, attn-gates, and projections, where $\theta = \{A_v, A_t, A_{attn}, G, P\}$.

4. Experiment

4.1. Dataset

CC595K[35] is filtered by leveraging a frequency-based filtering of terms, and a trimmed corpus consisting of 595k image-text pairs remains. We use CC595K to pre-train our model.

ScienceQA[50] consists of 21,000 data samples, including multiple-choice questions with multimodal content. It covers 3 subjects, 26 topics, 127 categories and 379 skills. We utilize the training and test split of the ScienceQA dataset to conduct evaluation of our model's performance.

LLaVA-mix665k[36] aggregates multiple vision-and-language datasets and systematically refines them into an instruction-based format suitable. We leverage the dataset to train models with the ability to perform diverse multimodal tasks. We conduct a capability assessment of our approach using the MME benchmark[51].

4.2. Implementation Details

For the image feature extraction part, we utilize the Q-former module from BLIP2 [10]. For the LLM, we select the LLaMA-2-chat [2] model. The default setting for the intermediate hidden layer dimension in our adapter is 32. We employ AdamW as the optimizer with a cosine learning rate decay strategy.

Method	#T-Param	Subject			Context Modality			Grade		Average
		NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	
Zero- & few-shot methods										
Human [50]	-	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
GPT-3.5 [50]	-	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97
GPT-3.5 (CoT) [50]	-	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
GPT-4 [3]	-	84.06	73.45	87.36	81.87	70.75	90.73	84.69	79.10	82.69
Representative & SoTA models										
MM-CoT$_{Large}$ [52]	738M	95.91	82.00	90.82	95.26	88.80	92.89	92.44	90.31	91.68
LLaVA [35]	13B	90.36	95.95	88.00	89.49	88.00	90.66	90.93	90.90	90.92
LLaVA+GPT-4(judge) [35]	13B	91.56	96.74	91.09	90.62	88.99	93.52	92.73	92.16	92.53
Parameter-efficient methods										
LLaMA-Adapter [7]	1.8M	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
LaVIN [8]	5.4M	89.88	94.49	89.82	88.95	87.61	91.85	91.45	89.72	90.83
PILL (ours)	43M	90.76	95.84	88.73	90.27	**89.49**	90.66	92.36	89.39	**91.30**

Table 1: Comparison on ScienceQA test set. Question classes: NAT = natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12. #T-Params denotes the number of trainable parameters.

Under our experimental settings, the total number of trainable parameters amounts to 43M. The training of our model is conducted on a single A6000. which is a hardware configuration widely acceptable among researchers.

4.3. Experimental Results

4.3.1. Results on ScienceQA

Overall performance. As shown in Table 1, we compare the performance of several existing methods on the ScienceQA [50]. Firstly, we consider zero-shot or few-shot approaches. Human performance in question answering achieves an accuracy of 88.40%. Remarkably, GPT-4 achieves an accuracy of 82.69% without any specific training on ScienceQA. Among other LLMs, MM-CoT$_{Large}$ [52] achieves an accuracy of 91.68%. We speculate that this performance difference may be due to variations in the base models. Notably, there is a significant discrepancy in the NAT and SOC metrics between MM-COT and LLaMA-based models. Among the LLM approaches built upon LLaMA, the fully fine-tuned LLaVA achieves performance comparable to MM-COT.

Settings	IMG.	Avg.
$\text{PILL}_{w/oMoMAE}$	85.62	89.04
$\text{PILL}_{w/oMAG}$	86.76	90.29
$\text{PILL}_{w/oSwiGLU}$	86.07	89.60
PILL	87.16	90.24

Table 2: Ablation studies on ScienceQA test set by using different modules in PILL

It is worth noting that these methods focus on full-model fine-tuning, resulting in high training costs. Furthermore, LLaVA leverages GPT-4 to assist in answer discrimination. Excluding these auxiliary methods, among the approaches that employ parameter-efficient strategies, LLaMA-Adapter and LaVIN exhibit a slight performance decrement compared to LLaVA. Compared to these approaches, our method achieves comparable performance to the state-of-the-art methods, even without employing any extravagant techniques. Our approach incorporates dedicated processing for the visual modality, consequently outperforming all other methods on the IMG metrics. In terms of parameter count, our method significantly reduces training overhead compared to LLaVA as we only need to train the adapter components of the model. Despite our method having significantly more training parameters compared to LaVIN, our training speed is faster due to the absence of gradient backpropagation to ViT during training. In fact, under the same experimental settings, our method requires only 90% of the training time compared to LaVIN. Overall, these results validate the efficiency and design of PILL.

Ablation study. To gain deeper insights into the impact of the proposed modules on model performance, we conduct the ablation experiments, and the results are reported in Table 2. It can be observed that with the MoMAE module removed, the overall performance of the model significantly decreased. After removing the MAG module, although the model's performance slightly improved, the absence of the MAG module influence the IMG score, adversely affecting the model's understanding of images.

Settings	IMG.	Avg.
Adapter$_{None}$	86.07	89.60
Adapter$_{GeLU}$	86.56	89.60
Adapter$_{SwiGLU}$	87.16	90.24

Table 3: Ablation studies on ScienceQA test set by using different activation function

Settings	w/o pre-train	w pre-train
LLaMA-Adapter [7]	85.19	-
LLaVA [35]	85.81	89.84
LaVIN [8]	89.41	-
PILL(ours)	**90.24**	**91.30**

Table 4: Results of existing multimodal LLMs in 7B model scaling on ScienceQA test set.

Removing the SwiGLU module also results in a decline in model performance. Overall, this set of ablation experiments clearly demonstrates the significant contributions of the proposed modules, MoMAE, MAG, and SwiGLU, to model performance.

In addition, to investigate the impact of different activation functions on the Adapter, we initially apply the None, GELU and SwiGLU activation functions. As is shown in Table 3, compared to None activation, GeLU yields marginal improvements on the IMG metric, it does not contribute to an overall performance gain. However, when we replace the activation function with our proposed SwiGLU activation function, we achieve a further improvement of 0.64%.

Effectiveness of pre-training. In Table 4, we compare the performance of PILL with other multimodal methods using the LLaMA-7B model specification. Without pre-training, LLaMA-Adapter[7], which was one of the pioneering methods to employ

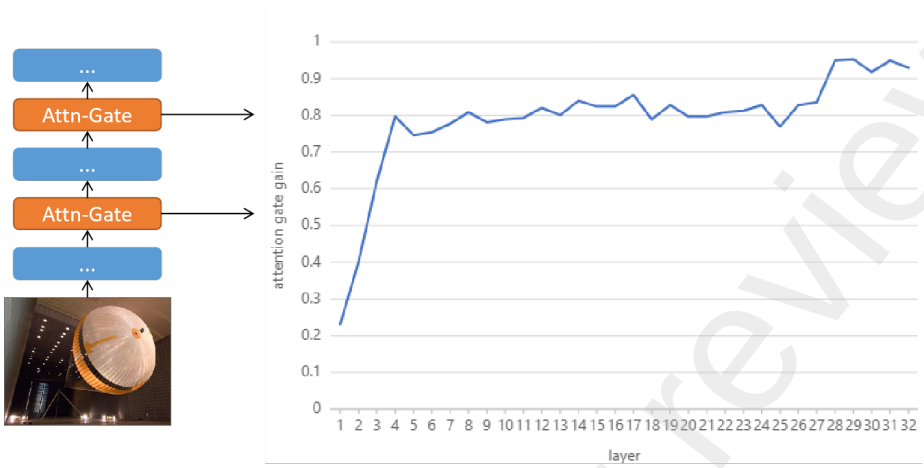

Figure 4: Evalution of the absolute value of the attention gate at different layers of PILL

PEFT techniques, enabled the LLM to possess visual recognition capabilities and achieved impressive results on ScienceQA[50] at that time. LAVIN[8] further improved upon this performance by approximately 4%. We speculate that the Adapter structure may be particularly suitable for VL-LLM, as evidenced by the experimental results of VL-Adapter[38]. On the other hand, since LLaVA[35] requires full fine-tuning, a well-initialized visual projection layer is likely to stabilize the subsequent fine-tuning process. As a result, LLaVA achieves a 4% improvement with pre-training. Importantly, our method consistently achieves the best results regardless of whether pre-training is employed, demonstrating the effectiveness of our proposed approach.

In-depth analysis on MAG layers. Figure 4 shows the absolute values of the MAG at different layers of our PILL model. We observe that the absolute values tend to increase with depth. We attribute this to the roles played by the LLM at different layers. In the layers closer to the input, the LLM focuses on language modeling and does not heavily rely on visual information. However, as the layers get closer to the output, the LLM has accumulated sufficient knowledge and needs to extract information from the images. As a result, the absolute values of the tanh gate activations start from near-zero values near the input layers and rapidly increase.

Impact of LLM hallucination on PILL. We explore the effect of LLM hallucination

Settings	IMG.	Avg.
w/o pre-train		
PILL	87.16	90.24
PILL$_{wrong}$	87.41	90.57
pre-train epoch=1		
PILL	88.25	90.73
PILL$_{wrong}$	87.85	90.64

Table 5: Hallucination studies on ScienceQA test set by setting wrong answer

on our model. In Table 5, under the setting without pre-training, we randomly replace the answer with another option with a probability of 0.1, leading to an additional improvement of 0.33%. We also explore the use of random wrong answers but encounter a decline in performance after we pre-train the model. We believe this is due to the lack of pre-training, which could result in insufficient image recognition capabilities and lead to hallucinations, similar to the findings in MM-COT[52]. After pre-training, the model exhibits improved image understanding, partially mitigating the hallucination issue. Therefore, employing the random wrong answer approach at this stage resulted in a decline in performance. Finally, by increasing the number of pre-training epochs to 3, we achieve the best results.

4.3.2. Qualitative Results

To provide more intuitive demonstrations that validate the effectiveness of our proposed method, we present several visualized examples for qualitative analysis. We highlight the model's incorrect outputs in red, while the correct output sentences are highlighted in green. As illustrated in Figure 5, When asked about the shared attributes in the image, our PILL model consistently provided the correct answer. However, the LaVIN model gave an inappropriate answer, incorrectly stating the existence of

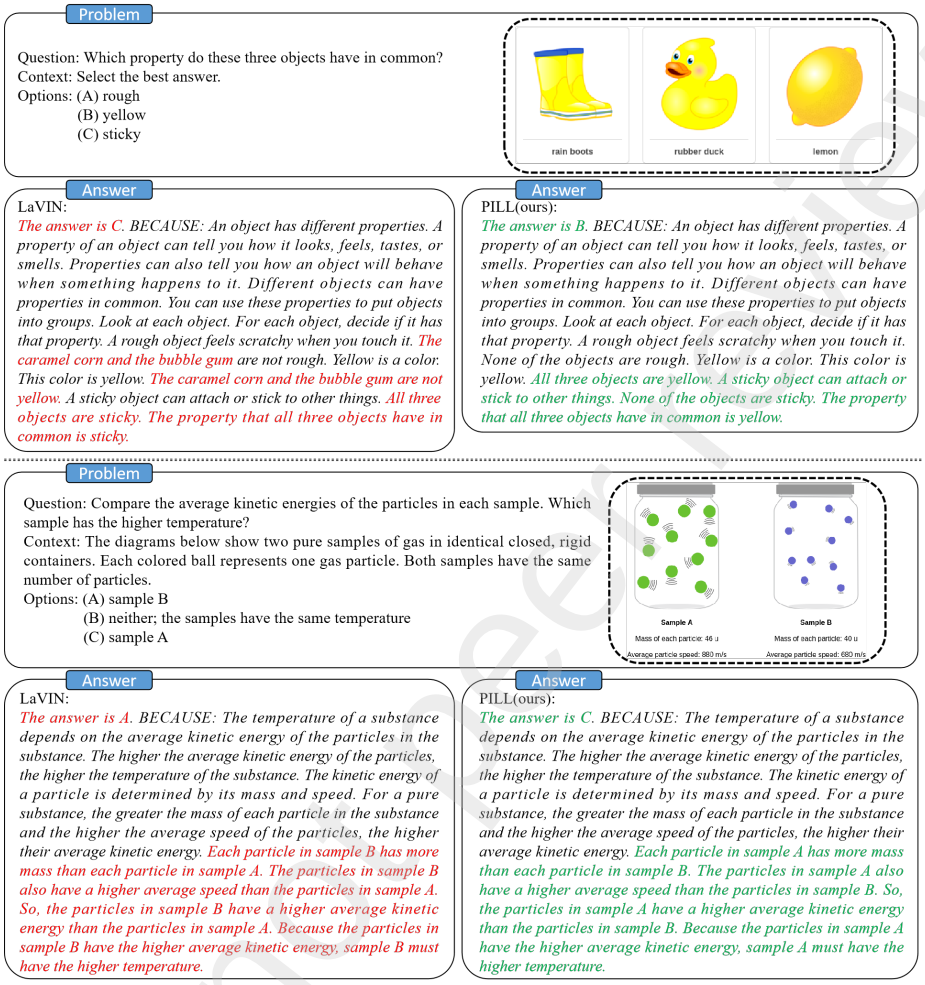

Figure 5: A comparison of PILL and LaVIN on ScienceQA test set

"the caramel corn and the bubble gum", which consequently led to an erroneous final response. Regarding the second question, the LaVIN model failed to accurately extract the key information of "mass" and "speed" from the image, resulting in an erroneous response. In contrast, our proposed PILL model successfully acquired these important visual cues from the image content.

Method	#T-Param	PT	FT	Perception	Cognition
Multimodal-GPT[41]	22.4M	0	288K	654.7	226.8
LaVIN[8]	5.4M	0	210k	963.6	249.6
LLaMA-Adapter-V2.1[47]	98.3M	200M	303K	1328.4	356.4
PILL (ours)	43.0M	595K	665K	1393.4	233.9

Table 6: Comparison of Parameter-efficient methods and quantitative evaluation results on MME benchmark

4.3.3. Results in MME Benchmark

We evaluate the generalization of our model's multimodal instruction following ability using the challenging MME benchmark. As shown in Table 6, previous state-of-the-art PEFT methods such as Multimodal-GPT and LaVIN achieved promising results without any pre-train. By further leveraging massive image-text pairs and multimodal instruction data, LLaMA-AdapterV2 established new standards with perceptive and cognitive scores of 1328.4 and 356.4 respectively through training more parameters. In contrast, our approach attains competitive metrics of 1393.4 and 233.9, demonstrating stronger performance gains despite considerably lower parameter budgets and training data requirements. This highlights the sample efficiency and parameter factorization abilities of our proposed method.

4.3.4. Open Demonstration

In Figure 6, we explore PILL's capability in zero-shot setting. We randomly selected two images from the COCO validation split and posed corresponding questions to PILL and LaVIN. In the first image, both PILL and LaVIN correctly output the man surfing. In the second instance, PILL accurately outputs the sailboat on the water, while LaVIN hallucinates by outputting non-existent entities such as other people and a kite. In the second image, both PILL and LaVIN correctly output the woman with an umbrella. However, PILL erroneously outputs the positional relationship between the two vehicles, while LaVIN erroneously outputs the woman's action, the color of her jacket and non-existent pedestrians in the image, and incorrectly mentions the presence

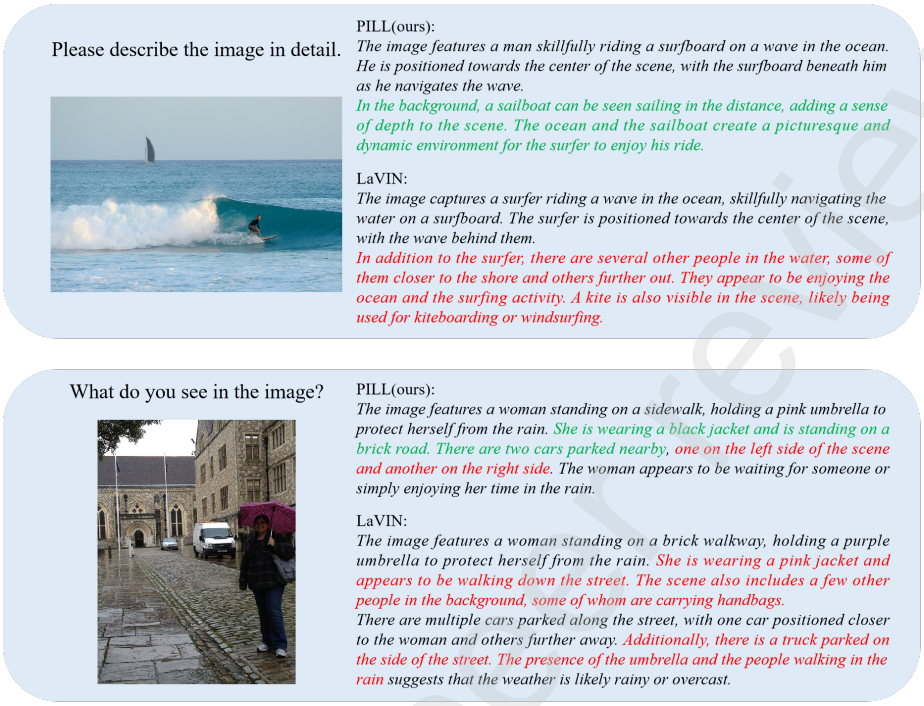

Figure 6: A comparison of PILL and LaVIN on Multimodal instructions following

of a truck.

5. Limitation and Future Work

Due to our utilization of LLM as the underlying model, PILL inevitably inherits some of the limitations of LLM. As a result, the image-to-text generation process may yield unsatisfactory results, including erroneous knowledge and reasoning or hallucinations. Despite the promising results achieved by PILL, we recognize that the model's ability to comprehend and leverage spatial relationships within visual scenes remains a challenge. To tackle this limitation, we plan to explore the incorporation of v-adapters that are specifically tailored to capture and process spatial and positional information more effectively. Additionally, although our work only involves the fusion of the image modality, our method can naturally incorporate other modalities such as video. In future work, we aspire to create a genuine multimodal AI assistant by further

exploring the integration of multiple modalities and scaling up the LLM.

6. Conclusion

In this work, we propose the PILL method to address the challenge of decoupling complex multimodal interactions. We leverage the Mixture-of-Modality-Adapter-Expert (MoMAE) module, which is specifically designed to handle image tokens, and the Modality-Attention-Gating (MAG) module for dynamic fusion of modalities. Additionally, we introduce the SwiGLU-Adapter to further enhance performance. Our experimental results demonstrate the effectiveness of our proposed method. With the advantage of fine-tuning only a small number of parameters, our approach offers a cost-effective solution that can be trained on a single A6000 GPU. This enables us to achieve visual language instruction following capabilities while maintaining efficient computation and training speed.

Acknowledgements

This work is supported by National Science and Technology Major Project (2022ZD0116700).

References

[1] OpenAI, Chatgpt (2023).

[2] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288.

[3] OpenAI, Gpt-4 technical report (2023).

[4] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al., Palm: Scaling language modeling with pathways, arXiv preprint arXiv:2204.02311.

[5] C. Li, H. Xu, J. Tian, W. Wang, M. Yan, B. Bi, J. Ye, H. Chen, G. Xu, Z. Cao, et al., mplug: Effective and efficient vision-language learning by cross-modal skip-connections, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 7241–7259.

[6] H. Xu, Q. Ye, M. Yan, Y. Shi, J. Ye, Y. Xu, C. Li, B. Bi, Q. Qian, W. Wang, et al., mplug-2: A modularized multi-modal foundation model across text, image and video, in: International Conference on Machine Learning, PMLR, 2023, pp. 38728–38748.

[7] R. Zhang, J. Han, C. Liu, A. Zhou, P. Lu, H. Li, P. Gao, Y. Qiao, LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention, in: The Twelfth International Conference on Learning Representations, 2024.

[8] G. Luo, Y. Zhou, T. Ren, S. Chen, X. Sun, R. Ji, Cheap and quick: Efficient vision-language instruction tuning for large language models, in: Advances in Neural Information Processing Systems, Vol. 36, 2023, pp. 29615–29627.

[9] N. Shazeer, Glu variants improve transformer, arXiv preprint arXiv:2002.05202.

[10] J. Li, D. Li, S. Savarese, S. Hoi, BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models, in: International Conference on Machine Learning, Vol. 202, PMLR, 2023, pp. 19730–19742.

[11] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, J. Dai, Vl-bert: Pre-training of generic visual-linguistic representations, in: International Conference on Learning Representations, 2020.

[12] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in: Advances in Neural Information Processing Systems, 2019, pp. 13–23.

[13] Z. Yu, J. Yu, Y. Cui, D. Tao, Q. Tian, Deep modular co-attention networks for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6281–6290.

[14] H. Tan, M. Bansal, Lxmert: Learning cross-modality encoder representations from transformers, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 5100–5111.

[15] F. Yu, J. Tang, W. Yin, Y. Sun, H. Tian, H. Wu, H. Wang, Ernie-vil: Knowledge enhanced vision-language representations through scene graphs, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 3208–3216.

[16] Y. Du, Y. Liu, Z. Peng, X. Jin, Gated attention fusion network for multimodal sentiment classification, Knowledge-Based Systems 240 (2022) 108107.

[17] Y. Cui, Z. Yu, C. Wang, Z. Zhao, J. Zhang, M. Wang, J. Yu, Rosita: Enhancing vision-and-language semantic alignments via cross- and intra-modal knowledge integration, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, Vol. 30, 2017.

[19] C. Chen, D. Han, X. Shen, Clvin: Complete language-vision interaction network for visual question answering, Knowledge-Based Systems 275 (2023) 110706.

[20] D. Song, S. Ma, Z. Sun, S. Yang, L. Liao, Kvl-bert: Knowledge enhanced visual-and-linguistic bert for visual commonsense reasoning, Knowledge-Based Systems 230 (2021) 107408.

[21] Y. Liu, X. Zhang, F. Huang, Z. Zhou, Z. Zhao, Z. Li, Visual question answering via combining inferential attention and semantic space mapping, Knowledge-Based Systems 207 (2020) 106339.

[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[23] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, S. C. H. Hoi, Align before fuse: Vision and language representation learning with momentum distillation, Vol. 34, 2021, pp. 9694–9705.

[24] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International Conference on Machine Learning, PMLR, 2022, pp. 12888–12900.

[25] H. Bao, W. Wang, L. Dong, Q. Liu, O. K. Mohammed, K. Aggarwal, S. Som, S. Piao, F. Wei, Vlmo: Unified vision-language pre-training with mixture-of-modality-experts, in: Advances in Neural Information Processing Systems, Vol. 35, 2022, pp. 32897–32912.

[26] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, F. Wei, Image as a foreign language: BEiT pretraining for vision and vision-language tasks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19175–19186.

[27] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, Y. Wu, Coca: Contrastive captioners are image-text foundation models, arXiv preprint arXiv:2205.01917.

[28] R. Mokady, A. Hertz, A. H. Bermano, Clipcap: Clip prefix for image captioning, arXiv preprint arXiv:2111.09734.

[29] X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, et al., Pali: A jointly-scaled multilingual language-image model, arXiv preprint arXiv:2209.06794.

[30] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, B. Patra, Q. Liu, K. Aggarwal, Z. Chi, N. Bjorck, V. Chaudhary, S. Som, X. SONG, F. Wei, Language is not all you need: Aligning perception with language models, in: Advances in Neural Information Processing Systems, Vol. 36, 2023, pp. 72096–72109.

[31] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, Q. Ye, F. Wei, Grounding multimodal large language models to the world, in: The Twelfth International Conference on Learning Representations, 2024.

[32] T. Lv, Y. Huang, J. Chen, L. Cui, S. Ma, Y. Chang, S. Huang, W. Wang, L. Dong, W. Luo, et al., Kosmos-2.5: A multimodal literate model, arXiv preprint arXiv:2309.11419.

[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, International Conference on Learning Representations.

[34] D. Zhu, J. Chen, X. Shen, X. Li, M. Elhoseiny, MiniGPT-4: Enhancing vision-language understanding with advanced large language models, in: The Twelfth International Conference on Learning Representations, 2024.

[35] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, in: Advances in Neural Information Processing Systems, Vol. 36, 2023, pp. 34892–34916.

[36] H. Liu, C. Li, Y. Li, Y. J. Lee, Improved baselines with visual instruction tuning, arXiv preprint arXiv:2310.03744.

[37] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for nlp, in: International Conference on Machine Learning, PMLR, 2019, pp. 2790–2799.

[38] Y.-L. Sung, J. Cho, M. Bansal, Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5227–5237.

[39] C. Eichenberg, S. Black, S. Weinbach, L. Parcalabescu, A. Frank, Magma–multimodal augmentation of generative models through adapter-based finetuning, in: Findings of the Association for Computational Linguistics: EMNLP 2022, 2022, pp. 2416–2428.

[40] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022.

[41] T. Gong, C. Lyu, S. Zhang, Y. Wang, M. Zheng, Q. Zhao, K. Liu, W. Zhang, P. Luo, K. Chen, Multimodal-gpt: A vision and language model for dialogue with humans, arXiv preprint arXiv:2305.04790.

[42] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. a. Bińkowski, R. Barreira, O. Vinyals, A. Zisserman, K. Simonyan, Flamingo: a visual language model for few-shot learning, in: Advances in Neural Information Processing Systems, Vol. 35, 2022, pp. 23716–23736.

[43] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, J. Tang, Glm: General language model pretraining with autoregressive blank infilling, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 320–335.

[44] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, et al., mplug-owl: Modularization empowers large language models with multimodality, arXiv preprint arXiv:2304.14178.

[45] Z. Yin, J. Wang, J. Cao, Z. Shi, D. Liu, M. Li, X. Huang, Z. Wang, L. Sheng, L. BAI, J. Shao, W. Ouyang, Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark, in: Advances in Neural Information Processing Systems, Vol. 36, 2023, pp. 26650–26685.

[46] X. L. Li, P. Liang, Prefix-tuning: Optimizing continuous prompts for generation, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 4582–4597.

[47] P. Gao, J. Han, R. Zhang, Z. Lin, S. Geng, A. Zhou, W. Zhang, P. Lu, C. He, X. Yue, H. Li, Y. Qiao, Llama-adapter v2: Parameter-efficient visual instruction model, arXiv preprint arXiv:2304.15010.

[48] Y. Huang, J. Lin, C. Zhou, H. Yang, L. Huang, Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably), in: International Conference on Machine Learning, PMLR, 2022, pp. 9226–9259.

[49] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971.

[50] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, A. Kalyan, Learn to explain: Multimodal reasoning via thought chains for science question answering, in: Advances in Neural Information Processing Systems, Vol. 35, 2022, pp. 2507–2521.

[51] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, Z. Qiu, W. Lin, J. Yang, X. Zheng, K. Li, X. Sun, R. Ji, Mme: A comprehensive evaluation benchmark for multimodal large language models, arXiv preprint arXiv:2306.13394.

[52] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, A. Smola, Multimodal chain-of-thought reasoning in language models, arXiv preprint arXiv:2302.00923.