

# IAA: Inner-Adaptor Architecture Empowers Frozen Large Language Model with Multimodal Capabilities

Bin Wang\*, Chunyu Xie\*, Dawei Leng†, Yuhui Yin

360 AI Research  
{wangbin10, xiechunyu, lengdawei, yinyuhui}@360.cn

## Abstract

In the field of multimodal large language models (MLLMs), common methods typically involve unfreezing the language model during training to foster profound visual understanding. However, the fine-tuning of such models with vision-language data often leads to a diminution of their natural language processing (NLP) capabilities. To avoid this performance degradation, a straightforward solution is to freeze the language model while developing multimodal competencies. Unfortunately, previous works have not attained satisfactory outcomes. Building on the strategy of freezing the language model, we conduct thorough structural exploration and introduce the Inner-Adaptor Architecture (IAA). Specifically, the architecture incorporates multiple multimodal adaptors at varying depths within the large language model to facilitate direct interaction with the inherently text-oriented transformer layers, thereby enabling the frozen language model to acquire multimodal capabilities. Unlike previous approaches of freezing language models that require large-scale aligned data, our proposed architecture is able to achieve superior performance on small-scale datasets. We conduct extensive experiments to improve the general multimodal capabilities and visual grounding abilities of the MLLM. Our approach remarkably outperforms previous state-of-the-art methods across various vision-language benchmarks without sacrificing performance on NLP tasks. Code and models are available at <https://github.com/360CVGroup/Inner-Adaptor-Architecture>.

## Introduction

Large Language Models (LLMs) have made substantial progress in recent years, largely attributed to the technique of pre-training and instruction tuning. Building upon this foundation, visual instruction tuning has been proposed to evolve LLMs into Multimodal Large Language Models (MLLMs), thereby endowing them with the capability to interpret and comprehend visual signals (Cha et al. 2024). MLLMs (Liu et al. 2024b; Bai et al. 2023; Tong et al. 2024; Chen et al. 2024b; Xuan et al. 2024) prove beneficial in numerous tasks, such as transcribing the text within an image,

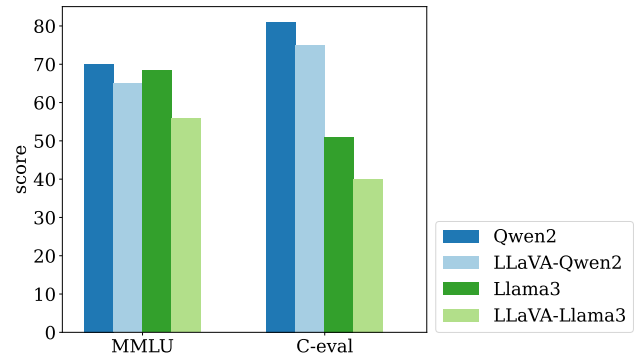


Figure 1: Results before and after training LLaVA-1.5 architecture based on Qwen2 and Llama3 language models on text-only evaluation set MMLU and C-eval.

generating stories and poems based on an image, or converting screenshots of webpages into code (Laurençon et al. 2024). Historically, these tasks have been regarded as challenging for conventional vision-language models. MLLMs exhibit considerable promise in executing these complex, diverse real-world tasks, enabling more natural and human-like interactions (Lu et al. 2024).

Typically, the operation of a MLLM begins with feeding an image into a visual encoder, such as CLIP (Radford et al. 2021) or SigLIP (Zhai et al. 2023), to extract a high-dimensional feature representation. This feature is subsequently transformed through a projection layer to align with the dimension of the large language model. The resulting features, often referred to as image tokens, are concatenated with text tokens and fed into the large language model. This process enables the MLLM to generate responses based on user instructions and input images.

In the current common MLLM (Liu et al. 2024a; Bai et al. 2023), when image and text tokens are fed into the large language model, the LLM is typically unfrozen for further training. This strategy has led to significant advancements in the MLLM model. Consequently, it predictably leads to a degradation in the understanding ability of the large language model. To validate this hypothesis, we conduct experiments on the LLaVA-1.5 (Liu et al. 2024a) architecture using the 1.2M-size open-source dataset provided by (Liu

\*These authors contributed equally.

†Corresponding author.

et al. 2024a), which contains a limited amount of plain text data, as illustrated in Figure 1. We compare the results before and after training the LLaVA-1.5 architecture, based on the Qwen2 (Yang et al. 2024) and Llama3 (Meta 2024) language models, respectively. The performance of the language model declines significantly on both the MMLU (Hendrycks et al. 2020) and C-Eval (Huang et al. 2023) text-only evaluation sets.

It appears reasonable to posit an explanation for this phenomenon within the field of deep learning. When a model is predominantly trained on a single type of data, it may experience a phenomenon known as catastrophic forgetting. For an MLLM to achieve outstanding image-text comprehension, it is essential to collect a substantial amount of image-text interaction data for training. As observed in Figure 1, training with image-text data results in a decline in language ability. Despite attempts by MLLM such as LLaVA to incorporate some text-only data into their training process, this still leads to a reduction in the model’s comprehension.

One direct method to prevent the degradation of LLM performance is to freeze the large language model during the training of MLLM. However, current methods employing this approach (Li et al. 2023a; Zhu et al. 2023) have consistently struggled to achieve powerful multimodal capabilities. To address these challenges, we propose a new training paradigm with an inner-adaptor architecture that significantly enhances multimodal competencies without affecting the original language modeling capabilities. This approach can seamlessly support both multimodal and textual workflows. We evaluate this training paradigm across a spectrum of tasks, including general multimodal capabilities and visual grounding proficiencies. Distinct from previous approaches of freezing language modeling that require large-scale aligned data, our proposed scheme demonstrates effectiveness with a considerably smaller dataset. Comprehensive testing on a suite of benchmarks, including MME, MMBench, MMMU, and RefCOCO, has substantiated the superior performance of our structure. We hope that this approach will provide a reference for future research in open-source MLLM.

## Related Work

**Large Language Models.** The landscape of Natural Language Processing (NLP) has undergone a revolutionary transformation, driven by the advent and continuous refinement of Large Language Models (LLMs). A pivotal moment in this evolution is the first appearance of the transformer architecture, which serves as a key catalyst, giving rise to pioneering language models like BERT (Devlin et al. 2018) and OPT (Zhang et al. 2022). These models showcase an unprecedented level of linguistic comprehension, significantly advancing the state-of-the-art in NLP. A critical breakthrough comes with introducing the Generative Pre-trained Transformer (GPT) series (Brown et al. 2020), which pioneer an auto-regressive language modeling approach, setting a new standard for language prediction and generation capabilities. Subsequent iterations, including Mixtral (Jiang et al. 2024), GPT-4 (Achiam et al. 2023), and Llama3 (Meta

2024), have not only maintained but also amplified this momentum, displaying superior performance on intricate language processing challenges. Moreover, the fusion of LLMs with specialized visual tasks showcases the models’ adaptability and broadens their scope, indicating their potential to transcend conventional text-based operations into multimodal interactions. This expansion highlights the transformative role LLMs can assume when incorporated into diverse domains, providing a rich ground for innovation and exploration.

**Multimodal Large Language Models.** The advancement of Large Language Models (LLMs) has kindled a growing interest in extending their foundational competencies to incorporate the visual domain, thereby giving birth to multimodal Large Language Models (MLLMs). The works on MLLMs (Xie et al. 2023; Li et al. 2023b,a; Bai et al. 2023; Liu et al. 2024b; Laurençon et al. 2024; Chen et al. 2024b) typically follow a tripartite architecture: a visual encoder, a vision-language connector, and a large language model. Notably, BLIP-2 (Li et al. 2023a) and Flamingo (Alayrac et al. 2022) introduce the Q-Former/Resampler as a bridge between vision and language, whereas LLaVA (Liu et al. 2024b) and MiniGPT4 (Zhu et al. 2023) refine this connection via a linear layer. Cambrian-1 (Tong et al. 2024) proposes a dynamically adaptive connector that integrates high-resolution visual features with LLMs while reducing the number of tokens. To enhance their multimodal performance, contemporary MLLMs mainly fine-tune the LLM and connector using visual instruction tuning data. These models leverage meticulously curated instruction datasets, showcasing an effective strategy that highlights their robust capabilities. However, a common oversight lies in the maintenance of language abilities. Long term multimodal training often leads to degradation of language proficiency. CogVLM (Wang et al. 2023) seeks to address this by integrating a trainable visual expert into the language model, but still trains the LLM during supervised fine-tuning, resulting in a degradation of language capability. DeekSeek-VL (Lu et al. 2024) maintains a 70% proportion of language data to preserve the integrity of language knowledge within the model, but incurs a considerable training cost. Departing from these conventional training paradigms of MLLMs, we introduce the inner-adaptor architecture. This design is specifically tailored to preserve the NLP performance of the MLLM while facilitating a seamless augmentation of its multimodal capabilities.

## Methodology

**Overview.** As illustrated in Figure 2, our approach enables the simultaneous execution of two high-quality workflows post-deployment: one for multimodal interactions and the other for text-only conversations. Both workflows leverage the transformer layers of the large language model. The multimodal interaction workflow encompasses: (1) an image encoder and a projector, utilized for extracting high-quality image features and achieving vision-language alignment, respectively, (2) the transformer layers of the large language model, which remain frozen during training, and (3) the

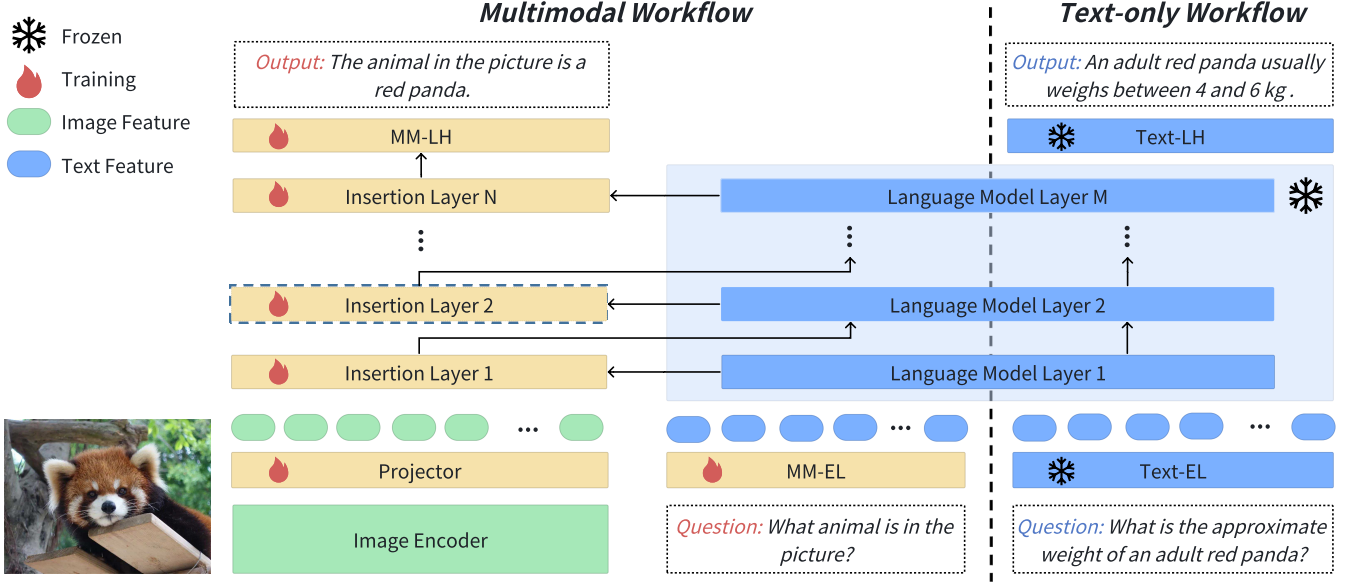


Figure 2: Overview of the proposed architecture, which mainly consists of two workflows: the Multimodal Workflow and the Text-only Workflow. The multimodal workflow, beyond the necessary image encoder and projector, integrates the Inner-Adaptor Architecture, including insertion layers, an embedding layer, and a language model head. Both workflows share the same large language model. The number of insertion layers is variable, where  $N \leq M$ . In this context, *MM* denotes MultiModal, *EL* stands for Embedding Layer, and *LH* represents the Language model Head.

inner-adaptor architecture, which comprises insertion layers, an embedding layer, and a language model head specifically designed for multimodal inputs. Conversely, the text-only conversation workflow solely employs the constituent elements of the original language model, without resorting to the specialized multimodal components.

**Image Encoder and Projector.** Following LLaVA-1.5 (Liu et al. 2024a), we utilize the CLIP ViT-L/14 (Radford et al. 2021) image encoder with an input resolution of 336px. Subsequently, we employ a vision-language projector composed of a two-layer MLP to integrate the vision features with LLMs.

**Large Language Model.** We employ the Llama3-8B (Meta 2024) as the base language model throughout the training process.

**Inner-Adaptor Architecture.** To achieve multimodal comprehension, it is essential to integrate trainable parameters into MLLMs. LLaVA (Liu et al. 2024b) makes the projector and the large language model trainable during visual instruction tuning, but leads to the performance degradation on NLP tasks. Flamingo (Alayrac et al. 2022) employs cross-attention with a gating mechanism to introduce image information into the model, facilitating a deep fusion of original image features with text features prior to each layer of the language model. However, this approach requires a considerable volume of pre-training data to train effective cross-attention layers and gating values, which can be computationally costly. Furthermore, the final performance of the model falls short of expectations.

Drawing insights from recent works (Chen et al. 2024a; Tong et al. 2024), we recognize that the self-attention layer can assimilate image features as prior prompts, thus eliminating the necessity of cross-attention for the obligatory incorporation of image features. In alignment with this perspective, we embark on exploratory research. Referencing Figure 3(a), we are inspired by the prevalent ControlNet (Zhang, Rao, and Agrawala 2023) architecture. The operation of a specific layer can be succinctly expressed as follows:

$$X_{out} = \phi_{fl}(X_{in}) + G(\phi_{il}(X_{in})), \quad (1)$$

where  $\phi_{fl}$  and  $\phi_{il}$  denote the frozen language model (LM) layer and the insertion layer, respectively. Here,  $X_{in}$  represents the multimodal input,  $X_{out}$  denotes the multimodal output, and  $G$  indicates a gating layer initialized at zero. The insertion layer is a transformer decoder layer, comprising the self-attention layer, layer normalization, feed forward network, etc. It is consistent with the parameter scale of a transformer layer in the large language model. For instance, if we target the 22<sup>th</sup> layer, the initial parameters of the corresponding insertion layer are derived from the 22<sup>th</sup> language model layer. Nonetheless, the ControlNet-based design did not yield satisfactory performance.

Referring to Figure 3(b), we endeavor to refine the ControlNet structure. Specifically, we eliminate the feature propagation between insertion layers. Instead, the output of the LM layer serves as the input to the insertion layer. Our expectation is that each frozen LM layer will accommodate multimodal data through a distinct insertion layer and gating layer, with the insertion layer no longer being directly

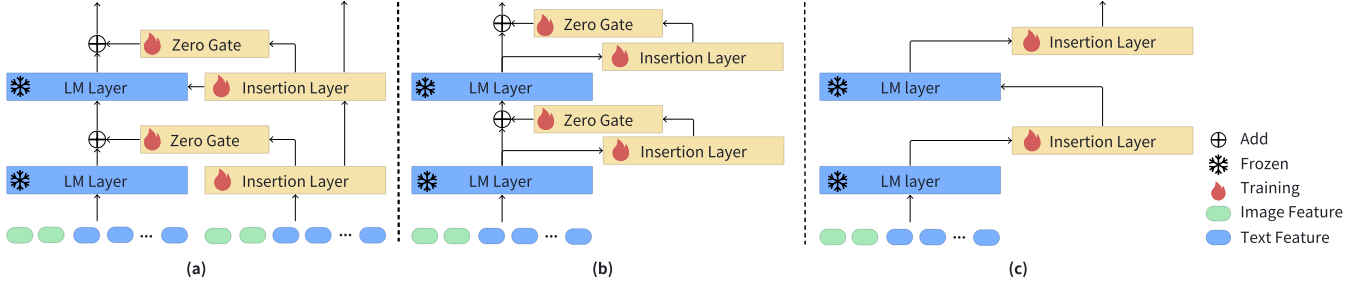


Figure 3: Structural exploration of the Inner-Adaptor Architecture. Figure (a) is a architecture inspired by the ControlNet design; Figure (b) is an improvement on Figure (a), mainly canceling the feature propagation between adaptors; Figure (c) is the final scheme.

influenced by subsequent layers. Compared to the design in Figure 3(a), the refined architecture shows significant improvements.

Moreover, we hypothesize that the gating layer may not reach an optimal state through a single round of data training. Consequently, we propose a more streamlined solution, as illustrated in Figure 3(c). The operation of a specific layer within the model can be represented as follows:

$$X_{out} = \phi_{il}(\phi_{fl}(X_{in})). \quad (2)$$

Similar to Scheme (a), if an insertion layer is placed after the 22th LM layer, it is initialized from the parameters of the 22th frozen LM layer. The number of insertion layers is adjustable.

Additionally, for multimodal training, we introduce a new embedding layer  $EL_{mm}$  and a new LM head  $LH_{mm}$ , initialized from the original language model’s embedding layer  $EL_{text}$  and LM head  $LH_{text}$ . Throughout all stages of multimodal training,  $EL_{text}$  and  $LH_{text}$  will remain frozen, while the newly created components will be trained with multimodal data. The experimental results presented in Table 5 validate the effectiveness of this strategy.

We thoroughly explore the distinctions among these architectures and strategies in the ablation study. Ultimately, we select the structure depicted in Figure 3(c), which we designate as the Inner-Adaptor Architecture (IAA).

## Experiments

In this section, we first describe the training paradigm of our method with the data utilized in the diverse processes. Subsequently, we conduct evaluation on the general multimodal and visual grounding benchmarks to comprehensively assess our models’ visual understanding ability. Finally, we detail the ablation experiments of our method.

### Training Paradigm

**Pre-training.** During the training process of MLLM, the primary objective of the pre-training phase is to enable MLLM to learn the alignment between visual cues and textual descriptions. This stage, also known as the image-text alignment phase, establishes connections between the vision encoder and LLM. In our architectural design, the image encoder and LLM remain frozen throughout all training phases

Configurations	Satge1-PT	Satge2-PT
Trainable modules	Projector	Projector, Inner-adaptor
Learning rate	1e-3	2e-5
Batch size		256
LR schedule		Cosine decay
Training steps		2.5K
Zero-Stage		Zero2
Warmup ratio		0.03
Weight decay		0.0
Optimizer		AdamW
Optimizer HPs	$\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1e - 6$	
Configurations	Instruction-FT	Grounding-FT
Trainable modules	Projector, Inner-adaptor	
Learning rate		2e-5
Batch size		128
LR schedule		Cosine decay
Training steps	6.6K	18K
Zero-Stage		Zero3
Warmup ratio		0.03
Weight decay		0.0
Optimizer		AdamW
Optimizer HPs	$\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1e - 6$	

Table 1: The hyperparameters utilized during the training phase are delineated as follows: ”-PT” designates the pre-training phase, ”-FT” denotes the fine-tuning phase, and ”HP” and ”LR” signify the hyperparameter and learning rate, respectively.

to preserve the inherent foundational knowledge in both vision and language models. The projector and inner-adaptor architecture require training to enhance multimodal capabilities. Our empirical investigations reveal that for the inner-adaptor architecture, applying a high learning rate can lead to overflow in training loss. To alleviate this issue, we devise a dual-stage pre-training procedure.

In the first pre-training stage, the model configuration consists of only three components: the image encoder, the projector, and the large language model. The parameters of the image encoder and the large language model are frozen, while a high learning rate of 0.001 is utilized to train a high-quality projector.

In the second pre-training stage, the model architecture

Method	Vision Encoder	Language Model	Data Scale	MME <sup>P</sup>	MMB-EN <sup>T</sup>	MMB-CN <sup>T</sup>	MMMU <sup>v</sup>
<i>Training with the LLM unfrozen</i>							
mPLUG-Owl(Ye et al. 2023)	CLIP-ViT-L	Llama2 (7B)	1.1B	967.3	49.4	-	-
Qwen-VL-Chat (Bai et al. 2023)	CLIP-ViT-G	Qwen (7B)	1.5B	1487.6	61.8	56.3	37
CogVLM (Wang et al. 2023)	EVA2-CLIP-ViT-E	Vicuna-v1.5 (7B)	1.5B	1439.7	65.8	55.9	37.3
mPLUG-Owl2 (Ye et al. 2024)	CLIP-ViT-L	Llama2 (7B)	400M	1450.2	66.0	60.3	34.7
LLaVA-1.5(Liu et al. 2024b)	CLIP-ViT-L	Vicuna-v1.5 (7B)	1.2M	1510.7	66.5	59.0	35.7
LLaVA-1.5(Liu et al. 2024b)	CLIP-ViT-L	Vicuna-v1.5 (13B)	1.2M	1531.3	69.2	65.0	37.0
Honeybee (Cha et al. 2024)	CLIP-ViT-L	Vicuna-v1.5 (7B)	208M	1584.2	70.1	-	-
Yi-VL (AI et al. 2024)	CLIP-ViT-H	Yi (6B)	125M	-	68.4	66.6	39.1
DeepSeek-VL (Lu et al. 2024)	SAM-B and SigLIP-L	DeepSeek (7B)	103M	-	73.8	<b>71.4</b>	36.6
LLaVA-Llama3 (Contributors 2024)	CLIP-ViT-L	Llama3 (8B)	1.2M	1506.0	68.9	61.6	36.8
<i>Training with the LLM frozen</i>							
OpenFlamingov2 (Awadalla et al. 2023)	CLIP-ViT-L	MPT (7B)	3B	-	5.7	14.4	28.8
Llama-AdapterV2(Gao et al. 2023)	CLIP-ViT-L	Llama2 (7B)	0.6M	972.7	41.0	-	-
MiniGPT-4 (Zhu et al. 2023)	EVA-CLIP-ViT-G	Vicuna (13B)	5.1M	866.6	-	-	-
BLIP-2 (Li et al. 2023a)	EVA-CLIP-ViT-G	FlanT5XXL	129M	1293.8	-	-	-
InstructBLIP (Dai et al. 2023)	EVA-CLIP-ViT-G	Vicuna (13B)	130M	1212.8	44.0	-	-
IAA-8 <sup>†</sup>	CLIP-ViT-L	Llama3 (8B)	1.2M	1560.2	69.9	64.2	39.0
IAA-8	CLIP-ViT-L	Llama3 (8B)	1.5M	1581.8	72.7	69.2	39.8
IAA-14	CLIP-ViT-L	Llama3 (8B)	1.5M	<b>1591.5</b>	<b>74.9</b>	70.5	<b>39.9</b>

Table 2: Results on general multimodal benchmarks, where the data scale of 1.2M uniformly represents the data provided by LLaVA (Liu et al. 2024b). IAA-8<sup>†</sup> represents the model trained using 1.2M data.

Method	MMLU <sup>†</sup>	C-Eval <sup>†</sup>	BBH <sup>†</sup>	Humaneval <sup>†</sup>	Math <sup>†</sup>
LLaVA-Llama3	55.8	40.5	44.6	38.4	12.3
IAA-8 <sup>†</sup>	68.4	51.3	52.8	59.2	27.8

Table 3: Comparison on Text-only Benchmarks. IAA-8<sup>†</sup> denotes the model trained using the same 1.2M data as LLaVA-Llama3. IAA-8<sup>†</sup> is not impaired in terms of NLP ability, but LLaVA-Llama3 presents deteriorated results.

is expanded to incorporate the inner-adaptor for multimodal tasks. The training parameters now include both the projector and the newly integrated structures. The projector is initialized with the parameters derived from the preceding stage. For this stage, a lower learning rate of  $2e-5$  is adopted.

Throughout the pre-training stages, the dataset employed consists of 558k image-text aligned pairs sourced from (Liu et al. 2024b) and an additional 100K pairs from (Chen et al. 2024a). (Chen et al. 2024a) provides a total of 664K image-text aligned data. We translate the first 100k pairs into Chinese and incorporated them into the training process to fortify the model’s understanding of Chinese tasks. Over the course of these stages, we utilize a cumulative total of 658K data pairs.

**Instruction Fine-tuning.** We perform instruction fine-tuning based on the model obtained from the second pre-training stage. Throughout this stage, the parameters of the large language model and the image encoder remain frozen. The dataset includes the fine-tuning dataset of 665K sam-

ples proposed by (Liu et al. 2024b), along with additional datasets including DocVQA (50K) (Mathew, Karatzas, and Jawahar 2021), VSR (10K) (Liu, Emerson, and Collier 2023), ScienceQA (21K) (Lu et al. 2022), and an in-house dataset (78.5K). Similar to the pre-training stage, we translate the first 40K entries of the 664K fine-tuning data proposed by (Chen et al. 2024a) into Chinese and incorporate them into the instruction fine-tuning dataset. The aggregate quantity of data utilized in this stage amounts to 865K.

**Grounding Fine-tuning.** Building upon the model fine-tuned with instructions, we further train a model specialized in visual grounding. The data used in this stage comprises RefCOCO (Kazemzadeh et al. 2014), COCO (Lin et al. 2014), Flickr30k Entities (Plummer et al. 2015), Objects365 (Shao et al. 2019), aggregating to approximately 2M data instances. These datasets improves the model’s capability of localizing fine-grained visual details. The inclusion of COCO and Objects365 assists the model in improving its ability to localize multiple targets.

**Implementation details.** The detailed training information is summarized in Table 1, mainly covering the hyper-parameters used during the four-stage training process. The entire four-stage can be executed on a single node A800 $\times$ 8 in 48 hours. All experiments utilize the zero technology provided by (Rajbhandari et al. 2020) and the flash-attention v2 provided by (Dao 2023).

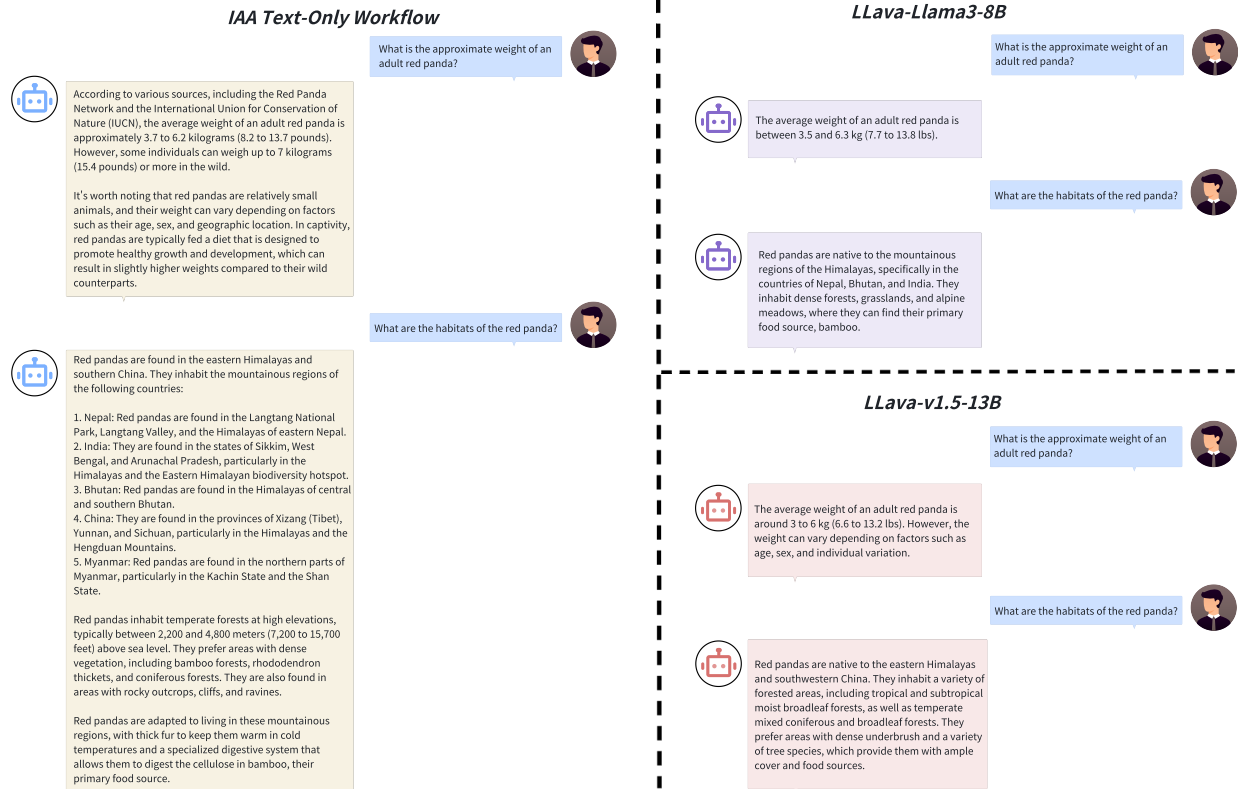


Figure 4: Comparison on text-only question answering.

## Experimental Results

**Main Results on General Multimodal Benchmarks.** To assess the multimodal capabilities of our approach, we employ widely recognized benchmarks that are closely related to multimodal tasks: MME<sup>P</sup> (Fu et al. 2023), MMBench-EN<sup>T</sup> (Liu et al. 2023), MMBench-CN<sup>T</sup> (Liu et al. 2023), and MMMU<sup>V</sup> (Yue et al. 2024). These benchmarks are renowned for presenting significant challenges across a diverse range of practical tasks. For evaluation purposes, we adhere to a zero-shot testing protocol, a strict methodology that tests models on unseen data without additional training. Moreover, we categorize comparative methods into two distinct categories: those trained with a frozen language model and those trained with an unfrozen language model. To provide a comprehensive analysis, we show the scale of the data utilized for each method, along with the variations in the image encoders employed. Detailed results of our evaluations are tabulated in Table 2. To ensure a fair and equitable comparison, we choose methods that leverage a base language model with a comparable parameter scale, and the reported metrics for competing methods are based solely on officially published data, avoiding any local testing results.

Owing to the inherent strengths of our proposed architecture, our method exhibits substantial superiority over those trained with frozen language model. As the current mainstream approach, models trained with unfrozen language models typically achieve better multimodal performance,

albeit at the cost of diminished NLP capabilities. We list several state-of-the-art methods adhering to this training paradigm. Compared to Honeybee (Cha et al. 2024), Yi-VL (AI et al. 2024), and Deepseek-VL (Lu et al. 2024), our method achieves competitive or even superior performance on certain metrics, with an extremely small training data scale. Using the same data scale of 1.2 million, IAA-8 outperforms LLaVA-Llama3. Additionally, IAA-14 with 14 insertion layers achieves better results than IAA-8 with an 8-layer configuration. Furthermore, we compare our approach with LLaVA-Llama3 (Contributors 2024) on NLP benchmarks, including MMLU and C-Eval. The results of NLP benchmarks are summarized in Table 3. Our language model is not impaired in terms of NLP ability, but LLaVA-Llama3 trained on the same data shows deteriorated results on both MMLU and C-Eval. Our method surpasses LLaVA-Llama3 across all metrics, indicating that our architecture is superior to the mainstream LLaVA architecture. The performance of various models on the plain text dialog task is illustrated in Figure 3. It is evident that the text-only workflow of the Inner-Adaptor Architecture (IAA) preserves the original conversational capabilities of the language model. In contrast, open-source multimodal large language models such as LLaVA-Llama3 and LLaVA-v1.5 are more impacted by multimodal data. When queried with the same question, LLaVA-Llama3 and LLaVA-v1.5 produce notably shorter responses. This is directly related to the fact that a



Method	Grounding Data Scale	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
KOSMOS-2 (Peng et al. 2023)	20M	52.3	57.4	47.3	45.5	50.7	42.2	60.6	61.7
OFA-L (Wang et al. 2022)	10M	80.0	83.7	76.4	68.3	76.0	61.8	67.6	67.6
Shikra (Chen et al. 2023b)	4M	87.0	90.6	80.2	81.6	87.4	72.1	82.3	82.2
MiniGPT-v2 (Chen et al. 2023a)	~21M	88.7	91.7	85.3	80.0	85.1	74.5	84.4	84.7
Ferret (You et al. 2023)	8.7M	87.5	91.4	82.5	80.1	87.4	73.1	83.9	84.8
PINK (Xuan et al. 2024)	5M	88.7	92.1	84.0	81.8	88.2	73.9	83.9	84.3
IAA-8	2M	89.2	92.6	83.7	82.1	88.6	73.7	84.4	84.7
IAA-14	2M	<b>90.2</b>	<b>92.9</b>	<b>85.4</b>	<b>83.4</b>	<b>89.0</b>	<b>76.7</b>	<b>85.0</b>	<b>85.1</b>

Table 4: Comparisons on visual grounding benchmarks. Our approach achieves competitive performance trained on relatively limited datasets.

large amount of the multimodal training data has shorter text lengths. Fine-tuning the large language model affects its ability to fully understand content and generate more comprehensive responses.

**Results on Visual Grounding Benchmarks.** To evaluate the effectiveness of our model in the visual grounding task, we perform evaluations utilizing the widely accepted benchmarks RefCOCO (Kazemzadeh et al. 2014), RefCOCO+ (Yu et al. 2016), and RefCOCOg (Mao et al. 2016), with the corresponding results illustrated in Table 4. The methods for comparison are all models trained for the grounding task under an auto-regressive strategy. The results reveal that our method is capable of achieving competitive performance, even when trained on relatively limited datasets. In our analysis, to ensure fairness, we exclude models trained on extremely large-scale datasets, such as CogVLM-grounding (Wang et al. 2023) with 1.5B image-text pairs and 40M grounding data, as well as those leveraging pre-trained object detection models, exemplified by LLaVA-Grounding (Zhang et al. 2023) and Groma (Ma et al. 2024).

**Efficiency in Deployment.** Currently, high-performance multimodal models typically require the unfreezing of the large language model for training. CogVLM (Wang et al. 2023) highlights the substantial difficulty in developing a model that excels in both multimodal comprehension and visual grounding tasks simultaneously. To address this, it adopts a dual-model strategy, specifically training one model for general multimodal capabilities and another for visual grounding abilities. In this context, deploying a high-quality language model, a multimodal model with outstanding general performance, and a model endowed with proficient visual grounding skills concurrently on a single GPU would demand an estimated 50GB of memory. Our proposed approach, facilitated by the inner-adaptor architecture, ingeniously combines superior general multimodal competencies and robust visual grounding capacities, while concurrently safeguarding the inherent prowess of the original large language model. Specifically, with an 8-layer inner-adaptor configuration, our model exhibits a significantly reduced

memory footprint, hovering around 30GB.

## Ablation Study

**Structure Analysis.** In the exploration of the structure, we furnish quantitative results for validation in Table 5. With an 8-layer insertion scheme as our baseline configuration, we observe that incremental architectural enhancements consistently improve performance metrics across the board. Specifically, the comparison between rows 1, 2, and 4 highlights the benefits of architectural refinement. Moreover, the contrast between rows 3 and 4 demonstrates that the integration of a specialized embedding layer and language model head for multimodal data processing significantly boosts performance.

**Comparison of Training Stages.** Through empirical evidence detailed in Table 6, we validate the effectiveness of our two-stage pre-training methodology. It can be observed that the model lacking the first stage of alignment training exhibits notably poorer performance. When the projector and insertion layers are engaged in joint pre-training, it is essential to maintain a learning rate of approximately  $2e-5$  to prevent loss overflow. However, this strategy leads to suboptimal alignment training for the projector, which negatively affects the model’s final performance. Furthermore, although the model performs adequately when skipping the second pre-training stage, it ultimately fails to replicate the outstanding results achievable through the complete two-stage pre-training process. This disparity emphasizes the critical significance of the additional pre-training stage in enhancing the model’s overall effectiveness.

**Impact of Insertion Layer Quantities.** We explore the effect of varying numbers of insertion layers, which are presented in Table 7. The experimental results indicate that increasing the number of insertion layers from 8 to 14 yields enhancements in all performance metrics. However, it is imperative to acknowledge that an increase in insertion layers simultaneously impacts the model’s efficiency. We advocate that an 8-layer configuration is adequate to effectively address foundational requirements.

Model architecture	Trainable modules					MME <sup>P</sup>	MMB-EN <sup>T</sup>	MMB-CN <sup>T</sup>	MMM <sup>U</sup>
	Projector	I-Layers(8)	$EL_{mm}$	$LH_{mm}$	Zero-Gates				
Figure 3(a)	✓	✓	✓	✓	✓	1425.4	72.4	65.0	38.2
Figure 3(b)	✓	✓	✓	✓	✓	1556.0	72.7	68.5	39.6
Figure 3(c)	✓	✓	×	×	×	1563.4	72.6	68.7	39.6
Figure 3(c)	✓	✓	✓	✓	×	1581.8	72.7	69.2	39.8

Table 5: Ablation study for the exploration of inner-adaptor related structures.

Training Stages			MME <sup>P</sup>	MMM <sup>U</sup>
Satge1-P	Satge2-P	Instruction-F		
×	✓	✓	1512.1	39.3
✓	×	✓	1565.4	39.5
✓	✓	✓	1581.8	39.8

Table 6: Comparison of the training stages.

Number of I-Layers	MME <sup>P</sup>	MMB-EN <sup>T</sup>	MMB-CN <sup>T</sup>
8	1581.8	72.7	69.2
14	1591.5	74.9	70.5
22	1531.7	76.0	70.4

Table 7: Ablations on the number of insertion layers.

**Training Data Influence Assessment.** To delineate the impact of data on model performance, we present comparative results in Table 8. The baseline, outlined in the first row, showcases the performance of LLaVA-Llama3 (Contributors 2024) utilizing the LLaVA architecture and the 1.2 million dataset provided by (Liu et al. 2024b). Subsequent experimentation, as delineated in the second row, emphasizes the pronounced superiority of our proposed architecture over LLaVA. Additionally, we enrich the training corpus with an extra 0.3 million records, mainly encompassing Chinese data. As a result, our model achieves substantial improvements in all metrics, especially on the Chinese evaluation set MMBench-CN<sup>T</sup>.

**Limitations** The method of extending multimodal capabilities by freezing the language model will introduce certain additional parameters. Compared to the approach of training with an unfrozen language model, the inference speed of the model will be reduced. To mitigate this issue, we extend the key-value cache mechanism to the insertion layers. Based on the MME dataset, compared to the LLaVA architecture, the average inference time of our 8-layer structure increases from 0.103s to 0.124s, which we deem to be within a relatively reasonable range.

## Conclusion

In this paper, we introduce the Inner-Adaptor Architecture, which is designed to enhance the general multimodal and visual grounding capabilities of LLMs. Through a series of architectural exploration experiments, we demonstrate that training with a frozen language model can surpass the mul-

	MME <sup>P</sup>	MMB-EN <sup>T</sup>	MMB-CN <sup>T</sup>
LLaVA-Llama3 (1.2M)	1506.0	68.9	61.6
IAA-8 (1.2M)	1560.2	69.9	64.2
IAA-8 (1.5M)	1581.8	72.7	69.2

Table 8: The impact of the training data.

timodal performance of the models with fine-tuned LLMs. Our proposed model has achieved state-of-the-art performance across a multitude of publicly available evaluation datasets. Moreover, after deployment, our approach incorporates dual workflows, thereby preserving the NLP proficiency of the language model. The flexibility of the Inner-Adaptor Architecture provides the potential for extension to additional modalities, which is a direction for future exploration.

## References

- Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- AI, .; ; Young, A.; Chen, B.; Li, C.; Huang, C.; Zhang, G.; Zhang, G.; Li, H.; Zhu, J.; Chen, J.; Chang, J.; Yu, K.; Liu, P.; Liu, Q.; Yue, S.; Yang, S.; Yang, S.; Yu, T.; Xie, W.; Huang, W.; Hu, X.; Ren, X.; Niu, X.; Nie, P.; Xu, Y.; Liu, Y.; Wang, Y.; Cai, Y.; Gu, Z.; Liu, Z.; and Dai, Z. 2024. Yi: Open Foundation Models by 01.AI. *arXiv:2403.04652*.
- Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in neural information processing systems*, volume 35, 23716–23736.
- Awadalla, A.; Gao, I.; Gardner, J.; Hessel, J.; Hanafy, Y.; Zhu, W.; Marathe, K.; Bitton, Y.; Gadre, S.; Sagawa, S.; Jitsev, J.; Kornblith, S.; Koh, P. W.; Ilharco, G.; Wortsman, M.; and Schmidt, L. 2023. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. *arXiv preprint arXiv:2308.01390*.
- Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.
- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners.



- In *Advances in neural information processing systems*, volume 33, 1877–1901.
- Cha, J.; Kang, W.; Mun, J.; and Roh, B. 2024. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13817–13827.
- Chen, G. H.; Chen, S.; Zhang, R.; Chen, J.; Wu, X.; Zhang, Z.; Chen, Z.; Li, J.; Wan, X.; and Wang, B. 2024a. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*.
- Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; and Elhoseiny, M. 2023a. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.
- Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023b. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*.
- Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.
- Contributors, X. 2024. XTuner: A Toolkit for Efficiently Fine-tuning LLM. <https://github.com/InternLM/xtuner>.
- Dai, W.; Li, J.; Li, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv preprint arXiv:2305.06500*.
- Dao, T. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; et al. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*.
- Gao, P.; Han, J.; Zhang, R.; Lin, Z.; Geng, S.; Zhou, A.; Zhang, W.; Lu, P.; He, C.; Yue, X.; et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.
- Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Huang, Y.; Bai, Y.; Zhu, Z.; Zhang, J.; Zhang, J.; Su, T.; Liu, J.; Lv, C.; Zhang, Y.; Lei, J.; Fu, Y.; Sun, M.; and He, J. 2023. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. In *Advances in Neural Information Processing Systems*.
- Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Hanna, E. B.; Bressand, F.; et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 787–798.
- Laurençon, H.; Tronchon, L.; Cord, M.; and Sanh, V. 2024. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.
- Li, J.; Xie, C.; Wu, X.; Wang, B.; and Leng, D. 2023b. What makes good open-vocabulary detector: A disassembling perspective. *arXiv preprint arXiv:2309.00227*.
- Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, 740–755. Springer.
- Liu, F.; Emerson, G.; and Collier, N. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11: 635–651.
- Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. In *Advances in neural information processing systems*, volume 36.
- Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2023. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.
- Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Sun, Y.; et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.
- Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521.
- Ma, C.; Jiang, Y.; Wu, J.; Yuan, Z.; and Qi, X. 2024. Groma: Localized Visual Tokenization for Grounding Multimodal Large Language Models. *arXiv preprint arXiv:2404.13013*.
- Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 11–20.
- Mathew, M.; Karatzas, D.; and Jawahar, C. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2200–2209.

- Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date. Technical report.
- Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.
- Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, 2641–2649.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Rajbhandari, S.; Rasley, J.; Ruwase, O.; and He, Y. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–16. IEEE.
- Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; and Sun, J. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8430–8439.
- Tong, S.; Brown, E.; Wu, P.; Woo, S.; Middepogu, M.; Akula, S. C.; Yang, J.; Yang, S.; Iyer, A.; Pan, X.; et al. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*.
- Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, 23318–23340. PMLR.
- Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.
- Xie, C.; Cai, H.; Li, J.; Kong, F.; Wu, X.; Song, J.; Morimitsu, H.; Yao, L.; Wang, D.; Zhang, X.; et al. 2023. CCMB: A Large-scale Chinese Cross-modal Benchmark. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4219–4227.
- Xuan, S.; Guo, Q.; Yang, M.; and Zhang, S. 2024. Pink: Unveiling the power of referential comprehension for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13838–13848.
- Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; Wang, J.; Yang, J.; Tu, J.; Zhang, J.; Ma, J.; Xu, J.; Zhou, J.; Bai, J.; He, J.; Lin, J.; Dang, K.; Lu, K.; Chen, K.; Yang, K.; Li, M.; Xue, M.; Ni, N.; Zhang, P.; Wang, P.; Peng, R.; Men, R.; Gao, R.; Lin, R.; Wang, S.; Bai, S.; Tan, S.; Zhu, T.; Li, T.; Liu, T.; Ge, W.; Deng, X.; Zhou, X.; Ren, X.; Zhang, X.; Wei, X.; Ren, X.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Fan, Z. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.
- Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; Jiang, C.; Li, C.; Xu, Y.; Chen, H.; Tian, J.; Qi, Q.; Zhang, J.; and Huang, F. 2023. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. *arXiv:2304.14178*.
- Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Hu, A.; Liu, H.; Qian, Q.; Zhang, J.; and Huang, F. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13040–13051.
- You, H.; Zhang, H.; Gan, Z.; Du, X.; Zhang, B.; Wang, Z.; Cao, L.; Chang, S.-F.; and Yang, Y. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*.
- Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, 69–85. Springer.
- Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9556–9567.
- Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11975–11986.
- Zhang, H.; Li, H.; Li, F.; Ren, T.; Zou, X.; Liu, S.; Huang, S.; Gao, J.; Zhang, L.; Li, C.; et al. 2023. Llava-grounding: Grounded visual chat with large multimodal models. *arXiv preprint arXiv:2312.02949*.
- Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *IEEE International Conference on Computer Vision (ICCV)*.
- Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.
- Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

## The Details of Datasets

In this section, we introduce the datasets IAA uses at different stages, along with the possible download links for these datasets in detail.

### Pre-training

Throughout the pre-training stages, the dataset employed consists of 558k image-text aligned pairs sourced from LLaVA and an additional 100K pairs from ALLaVA. ALLaVA provides a total of 664K image-text aligned data. We translate the first 100k pairs into Chinese and incorporated them into the training process to fortify the model's understanding of Chinese tasks. Over the course of these stages, we utilize a cumulative total of 658K data pairs.

**558k pairs from LLaVA** — <https://huggingface.co/datasets/liuhaotian/LLaVA-Pretrain>

<https://huggingface.co/datasets/liuhaotian/LLaVA-Pretrain>

**ALLaVA** — <https://huggingface.co/datasets/FreedomIntelligence/ALLaVA-4V>

### Instruction Fine-tuning

We perform instruction fine-tuning based on the model obtained from the second pre-training stage. Throughout this stage, the parameters of the large language model and the image encoder remain frozen. The dataset includes the fine-tuning dataset of 665K samples proposed by LLaVA, along with additional datasets including DocVQA (50K), VSR (10K), ScienceQA (21K), and an in-house dataset (78.5K). Similar to the pre-training stage, we translate the first 40K entries of the 664K fine-tuning data proposed by ALLaVA into Chinese and incorporate them into the instruction fine-tuning dataset. The aggregate quantity of data utilized in this stage amounts to 865K.

**665K samples from LLaVA** —

<https://huggingface.co/datasets/liuhaotian/LLaVA-Instruct-150K>

**DocVQA (50K)** —

<https://huggingface.co/datasets/cmarkea/doc-vqa>

**VSR (10K)** —

<https://github.com/cambridgeltl/visual-spatial-reasoning/>

**ScienceQA (21K)** —

<https://github.com/lupantech/ScienceQA>

**ALLaVA** — <https://huggingface.co/datasets/FreedomIntelligence/ALLaVA-4V>

<https://huggingface.co/datasets/FreedomIntelligence/ALLaVA-4V>

### Grounding Fine-tuning

Building upon the model fine-tuned with instructions, we further train a model specialized in visual grounding. The data used in this stage comprises RefCOCO, COCO, Flickr30k Entities, Objects365, aggregating to approximately 2M data instances. These datasets improves the model's capability of localizing fine-grained visual details. The inclusion of COCO and Objects365 assists the model in improving its ability to localize multiple targets.

**RefCOCO** — <https://github.com/lichengunc/refer>

**COCO** — <https://cocodataset.org/>

**Flickr30k Entities** —

[https://github.com/BryanPlummer/flickr30k\\_entities](https://github.com/BryanPlummer/flickr30k_entities)


**Objects365** — <https://www.objects365.org/>

## Supplementary Display

**Multimodal Capability.** Figures 5 and 6 showcase the capabilities of the Inner-Adaptor Architecture (IAA) in encyclopedia question answering, image comprehension, text recognition, and writing.

**Grounding Capability.** Figure 7 presents the multi-object detection capability of IAA, while Figure 8 demonstrates its detection capability for fine-grained perception.



What animal is in the picture?



The animal in the picture is a red panda.


What is the red panda doing in the picture?

The red panda is sitting on a wooden platform, possibly a tree stump, and looking at the camera.

(a)

Please explain the meme in detail.





The image features a small dog lying on its back on a wooden floor. The dog appears to be relaxed and comfortable in its position. The image is captioned with the words "Monday. Just. Monday." above the dog, and "Just... Monday" below it. The caption is a play on the phrase "Monday Blues," which is a common expression used to describe the feeling of being tired or unenthusiastic on a Monday. The image and caption together create a humorous and relatable representation of the feeling of dreading the start of the week.

In what other ways might someone express the same sentiment that this meme is expressing?

There are several ways someone might express the same sentiment as the "Monday Blues" meme:

1. A person might say, "Ugh, it's Monday again. I don't want to go to work/school."
2. They could post a picture of themselves looking tired or unenthusiastic on social media with the caption "Monday Blues."
3. A person might share a quote or saying related to the feeling of dreading the start of the week, such as "Monday morning: when you realize it's time to go back to work/school."
4. They could create a humorous skit or video that portrays the struggles of waking up on a Monday and trying to get motivated for the day ahead.
5. A person might even create a song or rap about the "Monday Blues," expressing their feelings of tiredness and reluctance to start the week.

These are just a few examples of how someone might express the same sentiment as the "Monday Blues" meme, but the key idea is to convey the feeling of dreading the start of the week and the desire to avoid or delay the responsibilities and tasks that come with it.






(b)

Figure 5: Samples of image comprehension and general knowledge question answering.

What is the sentence in the picture?

Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is.





The sentence in the picture is: "Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is."

Can you explain this meme in detail?

The image is a humorous representation of the Earth, created using chicken nuggets. The chicken nuggets are arranged to resemble the continents of the world, with each piece of chicken representing a specific region. The image is presented as a joke, as it is not possible to create an accurate representation of the Earth using only chicken nuggets. The meme is meant to be amusing and lighthearted, as it takes a common food item and gives it a unique and unexpected twist.

(a)

Write a fairy tale based on this painting.





Once upon a time, in a lush forest, there lived a bear, a cat, and a rabbit. They were the best of friends and loved spending time together. One day, they decided to have a picnic in the woods.

As they sat around the table, they noticed that the table was set with a variety of delicious food, including cookies, cakes, and other treats. The bear, being the largest of the three, was the first to dig in. He savored each bite of his meal, enjoying the flavors and textures of the food.

Next, the cat sat down and began to eat her share of the meal. She was particularly fond of the cookies and cakes, and she couldn't get enough of them. The rabbit, being the smallest of the three, was a bit more cautious in her eating, but she still enjoyed the food and the company of her friends.

As they ate, they chatted and laughed, sharing stories and jokes. The bear told a tale of his adventures in the forest, while the cat regaled them with stories of her mischievous escapades. The rabbit, being the quietest of the group, listened intently and added her own thoughts and ideas to the conversation.

After they finished their meal, the three friends decided to take a walk through the woods. They explored the forest, discovering new sights and sounds, and enjoying each other's company. As the sun began to set, they made their way back to their homes, already looking forward to their next adventure together.

And so, the bear, the cat, and the rabbit lived happily ever after, sharing their love

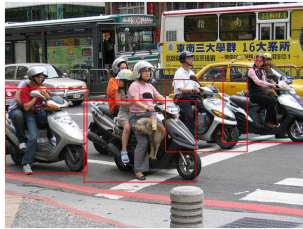
(b)

Figure 6: Samples of text recognition and writing ability.

Please provide the bounding box coordinate of the region this sentence describes:  
motorcycle.



[0.0, 0.41, 0.27, 0.69]; [0.26, 0.45, 0.63, 0.72]; [0.53, 0.38, 0.8, 0.62]; [0.72, 0.34, 1.0, 0.59]



Please provide the bounding box coordinate of the region this sentence describes:  
bicycle.



[0.32, 0.53, 0.92, 1.0]; [0.41, 0.52, 0.92, 0.99]; [0.26, 0.4, 0.42, 0.54]; [0.41, 0.39, 0.47, 0.5]; [0.45, 0.4, 0.51, 0.48]; [0.58, 0.38, 0.92, 0.98]

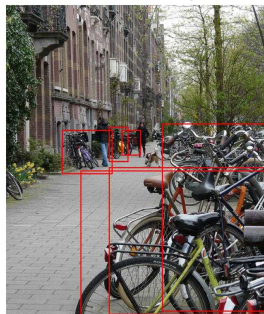


Figure 7: Samples of multi-object detection .



Please provide the bounding box coordinate of the region this sentence describes: dude with black shirt says circa.



[0.78,0.33,1.0,0.86]



Please provide the bounding box coordinate of the region this sentence describes: man in striped shirt holding wine glass.



[0.52,0.19,0.83,0.99]



Figure 8: Samples of fine-grained detection.