

A Comprehensive Overhaul of Multimodal Assistant with Small Language Models

Minjie Zhu^{1*}, Yichen Zhu^{2*†}, Ning Liu², Xin Liu¹, Zhiyuan Xu², Chaomin Shen^{1†}, Yaxin Peng³

¹ East China Normal University

² Midea Group

³ Shanghai University

{zhumj, liuxin}@stu.ecnu.edu.cn, {zhuyc25, xuzy70}@midea.com,
ningliu1220@gmail.com, cmshen@cs.ecnu.edu.cn, yaxin.peng@shu.edu.cn

Abstract

Multimodal Large Language Models (MLLMs) have showcased impressive skills in tasks related to visual understanding and reasoning. Yet, their widespread application faces obstacles due to the high computational demands during both the training and inference phases, restricting their use to a limited audience within the research and user communities. In this paper, we investigate the design aspects of Multimodal Small Language Models (MSLMs) and propose an efficient multimodal assistant named Mipha, which is designed to create synergy among various aspects: visual representation, language models, and optimization strategies. We show that without increasing the volume of training data, our Mipha-3B outperforms the state-of-the-art large MLLMs, especially LLaVA-1.5-13B, on multiple benchmarks. Through detailed discussion, we provide insights and guidelines for developing strong MSLMs that rival the capabilities of MLLMs.

Code — <https://github.com/zhuyiche/llava-phi>

Introduction

Recent advancements in Multimodal Large Language Models (MLLMs) have demonstrated exceptional visual understanding and reasoning performances across a range of tasks such as visual question answering (Team et al. 2023; Alayrac et al. 2022; Wang et al. 2023), referring comprehension (Achiam et al. 2023), and for embodied intelligence (Wen et al. 2024b; Zhu et al. 2024c; Wen et al. 2024a; Zhu et al. 2024b; Wen et al. 2024c; Zhu et al. 2024d). Benefiting from the scaling law of Transformer architecture (Vaswani et al. 2017) and web-scale training data sources, these models have become foundational in the field of artificial intelligence, with their parameters increasing from billions to trillions (Achiam et al. 2023; Touvron et al. 2023; Chowdhery et al. 2023). However, the deployment of these models is often hindered by their substantial computational costs and memory consumption in both the training and inference phases, which limits their popularization across the broader research and user communities.

*These authors contributed equally.

†: Corresponding authors.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

How can we boost the inference speed of MLLMs? Delving into the computational costs of MLLMs, it becomes evident that the Large Language Models (LLMs), primarily tasked with transforming visual-text pairs into textual outputs, account for a substantial computational load. In MLLMs, the visual encoder, such as CLIP-ViT-L, typically only comprises around 0.4 billion parameters, while the language model within MLLMs can escalate to between 7 and 65 billion parameters.

Intuitively, reducing the computational demands of the language model could lead to a significant decrease in overall inference costs. The rapid advancements by the open-source community in developing Small Language Models (SLMs), with parameter sizes ranging from 1 billion to 3 billion (i.e., Phi (Gunasekar et al. 2023; Li et al. 2023a; Javaheripi et al. 2023)), enable us to construct vision-language models more compact than traditional MLLMs. However, as the parameter count of these smaller language models is reduced, their capabilities often do not measure up to those of their larger counterparts, potentially diminishing the effectiveness of small multimodal models. This motivation drives us to explore the development of multimodal assistants using small language models that can deliver competitive performance comparable to their counterparts (Dai et al. 2023; Bai et al. 2023; Ye et al. 2023; Liu et al. 2023a), which rely on large language models.

In this paper, we present an in-depth analysis of the impact of various design elements in MSLMs on their downstream performance. A brief summary of the design space that we explored is illustrated in Fig 1 (right). Our investigation dissects the existing MSLMs into three key components: the language model, the visual representation, and the optimization strategy. Our thorough empirical research leads us to **several findings that diverge from the conventional wisdom** established by prior studies on Multimodal Large Language Models. For example:

- Increasing image resolution is not a silver bullet. In some benchmarks, images with a resolution of 224 pixels outperform those with 448 pixels.
- Contrary to earlier findings that suggest finetuning the visual backbone negatively impacts MLLMs, we underscore the importance of simultaneously finetuning both the visual backbone and the language model for MSLMs.
- While the prevailing trend among MLLMs is to employ

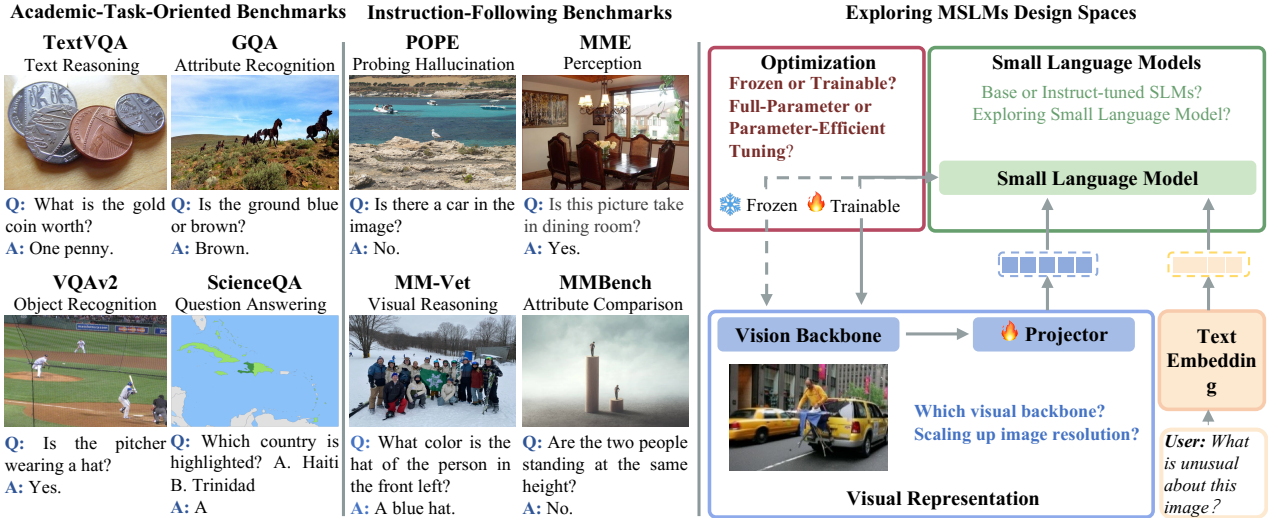


Figure 1: **Evaluation Benchmarks Overview.** We evaluate our model variants on academic-task-oriented benchmarks (left) as well as instruction-following benchmarks (middle). The answer is predicted by our proposed efficient MSLM: Mipha-3B. Additionally, we explore three key design spaces of MSLMs: 1) visual representation, 2) language model, and 3) optimization strategy (right).

instruct-tuned language models, such as Vicuna (Chiang et al. 2023), our analysis reveals that instruction tuning—be it through supervised fine-tuning or reinforcement learning from human feedback (RLHF)—is not essential.

Building on these findings, we present a new family of MSLMs named Mipha, scaled from 1.7B to 3B, which surpasses the performance of leading open-source MLLMs including LLaVA-1.5 and Qwen-VL, without the need for additional training data. Remarkably, our Mipha-3B models demonstrate superior performance across a majority of benchmarks when compared to the 7B MLLM variants, such as LLaVA-1.5, Qwen-VL, and InstructBLIP, and in some cases, even outperform 13B MLLM counterparts. Overall, we believe that our analysis provides a practical new perspective on optimizing effective MSLMs. Overall, our contributions are the follows:

1. We offer a detailed examination of the design landscape for Multimodal Small Language Models, examining visual components, language models, and optimization strategies from a comprehensive perspective. Our empirical research provides unique findings, **presenting multiple conclusions that contrast with prior studies on Multimodal Large Language Models.**
2. By integrating our empirical discoveries, we unveil a series of Multimodal Small Language Models named Mipha. This advancement underscores the practical impact of our findings.
3. Demonstrating the effectiveness of our approach, we show that our flagship model, Mipha-3B, outperforms existing 7B and 13B models on various benchmarks without requiring additional data for training. This marks a significant

advancement for multimodal models with fewer than 3B parameters, establishing a new benchmark in the field.

Preliminaries

In this section, we provide an overview of model architecture and evaluation benchmarks.

Overview of Model Architecture

The architecture of mainstream MLLMs (Liu et al. 2023a; Zhu et al. 2024a; Ye et al. 2023) typically comprises three primary components: 1) a visual representation backbone V_ψ , 2) a vision-language projector F_ω aimed at aligning visual and textual domains, and 3) a language model LM_θ .

Visual Representation Backbone. Given an input image X_{img} , we consider a visual representation backbone to output the visual features $V_{img} \in \mathbb{R}^{L \times h_v}$, where $V_{img} = V_\psi(X_{img})$. The L denotes the number of patches. All of our experiments consider only the grid features before and after the last transformer layer.

Vision-Language Projector. Then, a vision-language projector is employed to convert the visual features V_{img} to the text embedding space $T_{img} \in \mathbb{R}^{L \times h_t}$, where $T_{img} = F_\omega(V_{img})$. Following LLaVA-1.5 (Liu et al. 2023a), we employ a two-layer MLP as our projector across all of our model variants, which can effectively align visual and textual features.

Language Model (LM). Finally, based on T_{img} and the question embedding E_t , a language model is used to give the answer A , where $A = LM_\theta(\text{cat}(T_{img}, E_t))$. While previous studies mainly focus on language models with more than 7 billion parameters, such as Vicuna or LLaMA.

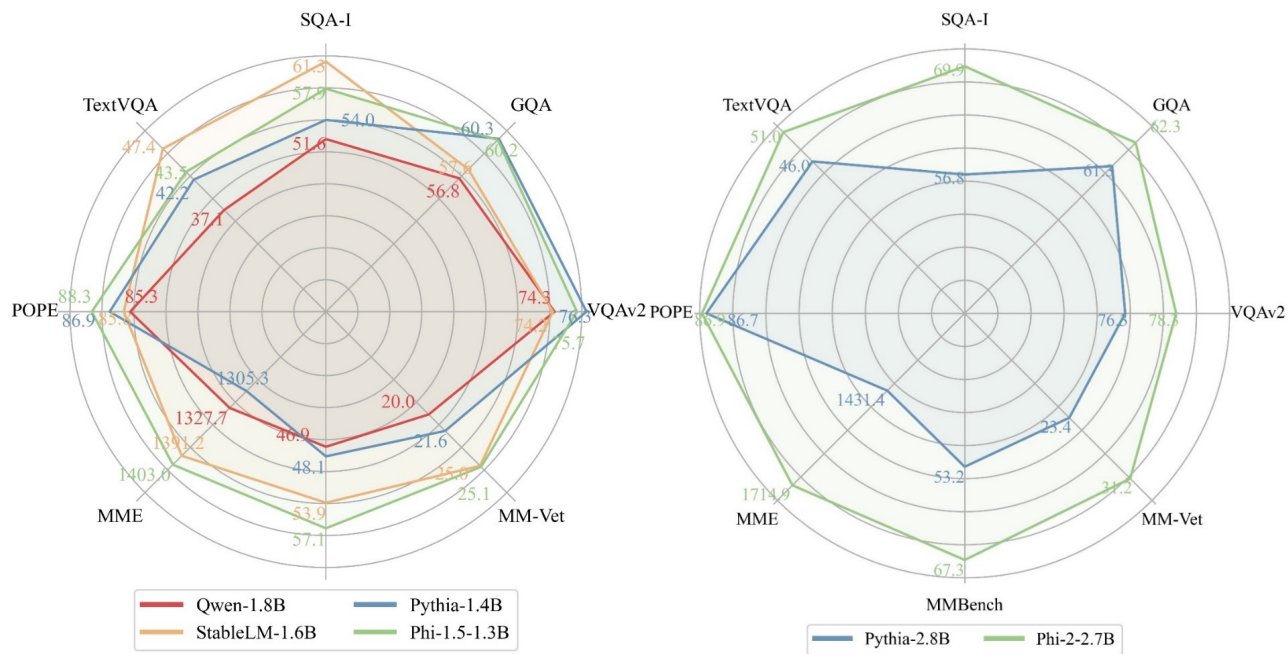


Figure 2: **Selection of Small Language Models.** We evaluate four open-sourced SLMs on 8 benchmarks, and Phi-family showcases the best performance (**left & right**).

Evaluation Benchmarks

There are many existing benchmarks (Li et al. 2024; Liu et al. 2025) for evaluation the capability of MLLMs. We conduct empirical studies on a collection of both academic-task-oriented benchmarks and recent benchmarks specifically proposed for instruction-following MLLMs, totaling 8 benchmarks. For academic-task-oriented benchmarks, VQA-v2 (Goyal et al. 2017) and GQA (Hudson and Manning 2019) evaluate the model’s visual perception capabilities on open-ended short answers. ScienceQA (Lu et al. 2022) is used to evaluate the zero-shot generalization on scientific question answering. TextVQA (Singh, Natarajan et al. 2019) contains text-rich visual question answering. We also employ recent benchmarks proposed for instruction-following MLLMs. POPE (Li et al. 2023b) evaluates MLLM’s degree of hallucination on three sampled subsets of COCO (Lin, Maire et al. 2014). MME (Fu et al. 2023) Benchmark evaluates MLLMs’ perception and cognition capabilities, and MMBench (Liu et al. 2023d) evaluates the model’s answer robustness with all-around shuffling on multiple-choice answers. MM-Vet (Yu et al. 2023) evaluates MLLM’s capabilities in engaging in visual conversations on a diverse range of tasks. Some examples are illustrated in Fig 1.

Exploring Crucial Aspects on Efficient MSLMs

There are a few prior studies that have explored different aspects of enhancing MLLMs (Karamcheti et al. 2024; Lu et al. 2023; Zhu et al. 2024e). However, these works primarily focus on multimodal assistants powered by large language models. Intuitively, the model insights obtained from multimodal large language model should also apply to multimodal

smaller language models. Nevertheless, in the following section, the contrary: techniques effective for large-scale models do not always seamlessly transfer to small-scale models. We also demonstrate several new observations that can improve the performance of MSLMs. We highlight that all our experiments in this section utilize the same amount of training data, and unless otherwise stated, all training components are maintained consistently for a fair comparison.

What Makes Small Language Model Good for Multimodal Assistant?

In this section, we delve into the methodology for selecting and training an effective Small Language Model (SLM) that enhances the performance of Multimodal Small Language Models (MSLMs). Our discussion is twofold. Initially, we evaluate various SLMs to determine which model exhibits optimal performance upon integration with a visual backbone. Subsequently, we investigate the MSLMs’ performance with base or instruct-tuned language models.

Which Small Language Model should we adopt? As the "brain" of multimodal models, the selection of pretrained small language models is crucial for crafting high-performing Multimodal Small Language Models (MSLMs). We evaluate a spectrum of recently introduced Small Language Models (SLMs) with sizes ranging from 1.3B to 2.8B parameters. This includes models such as Phi-2 (2.7B) (Javaheripi et al. 2023), Phi-1.5 (1.3B) (Li et al. 2023a), Qwen-1.5 (1.8B) (Bai et al. 2023), Pythia (available in both 1.4B and 2.8B sizes) (Biderman et al. 2023), and StableLM (1.6B) (Stability 2024). For analytical clarity, we classify these SLMs into two categories: those models possess over 2.7B parameters, and the

Method	MMLU	BBH	GSM8K	HumanEval
Pythia-1.4B	25.4	-	-	4.3
StableLM-1.6B	45.9	37.7	52.5	35.4
Qwen1.5-1.8B	46.8	24.2	38.4	20.1
Phi-1.5-1.3B	43.9	-	44.6	41.4
Pythia-2.8B	28.3	-	-	5.1
Phi-2-2.7B	56.7	43.4	57.2	47.6

Table 1: Comparison of small language models on Massive Multitask Language Understanding (MMLU) (Hendrycks et al. 2021), Instruction Following (BBH) (Srivastava et al. 2022), Math (GSM8K) (Cobbe et al. 2021), and Coding (HumanEval) (Chen et al. 2021).

remainder, feature less than 2B parameters. As illustrated in Figure 2 (left), we assess the performance of MSLMs that incorporate language models with fewer than 2 billion parameters. Phi-1.5-1.3B, despite having fewer parameters compared to its counterparts, still manages to display competitive or even superior performance. To delve into the reasons behind Phi-1.5’s outperformance, we benchmark these SLMs against four tests traditionally used for LLMs, as depicted in Table 1. These benchmarks include the MMLU for language understanding, BBH for instruction following, GSM8K for mathematics, and HumanEval for coding. Our observations reveal that Phi-1.5-1.3B remains competitive or outperforms models with a larger parameter count, such as Pythia-1.4B, StableLM-1.6B, and Qwen-1.8B. This suggests that Phi-1.5-1.3B makes efficient use of its parameters, particularly in the areas of reasoning and language understanding that these benchmarks measure, enhancing its transferability to multimodal contexts. Our findings indicate that a strong small language model can lead to better multimodal performance.

Moreover, when comparing Phi-2-2.7B with Pythia-2.8B, two relatively larger SLMs, the superiority of Phi-2 is evident. Aside from POPE, where models with Pythia show slightly fewer object hallucinations, Phi-2 models achieve markedly better outcomes on other metrics. This aligns with expectations, as Phi-2 was released after Pythia and was pretrained on a considerably larger dataset—1.4 trillion tokens versus 300 billion tokens. Additionally, the qualitative results corroborate our findings. MSLMs based on the Phi models provide detailed descriptions derived from the images. Notably, the MSLM equipped with Phi-2-2.7B goes a step further by recognizing that the monkey is performing, which showcases a robust capacity for scene reasoning. The examples and detailed analyses are provided in the appendix.

Base vs. Instruct-tuned LMs. The most state-of-the-art MLLMs (Liu et al. 2023a; Dai et al. 2023) leverage instruct-tuned LM, such as Vicuna, as the default backbone. However, employing instruction tuning on LMs means extra training time, and it also brings bias and regressions in performance (Ouyang et al. 2022). As such, we investigate whether instruction tuning is necessary for MSLMs. We consider the following four settings:

- **Base.** The base language model (LM) inherits the weights from pretrained models as is. Typically, this base LM ex-

hibits limited proficiency in following instructions, due to its training on a general corpus without specific emphasis on instruction adherence.

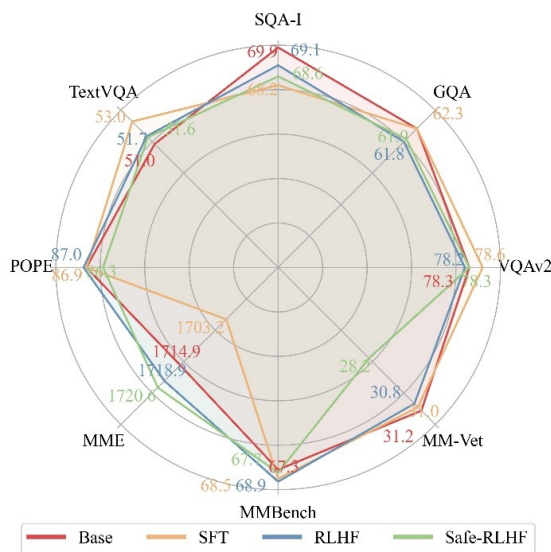
- **Supervised Fine-tuning (SFT).** SFT refines a pretrained model’s parameters using a targeted, often smaller dataset with labeled examples to enhance its performance for a specific task. Within the realm of Multimodal Large Language Models (MLLMs), employing a language model enhanced through SFT, such as Vicuna, serves as a prevalent strategy.
- **Reinforcement Learning from Human Feedback (RLHF).** RLHF (Ouyang et al. 2022) finetunes LMs by integrating preferences or feedback directly from humans. This method refines the capabilities of base LMs to function effectively as dialogue agents.
- **Safe Reinforcement Learning from Human Feedback (Safe-RLHF).** Safety alignment is a critical step to mitigate the risk of LMs generating undesirable outputs in response to certain prompts (Liu et al. 2023c).

The experimental findings are presented in Fig 3 (left), revealing that there is no universal solution to enhancing model generalization. The comparative analysis across all 8 benchmarks shows minimal variance among the four configurations tested. Notably, the MSLM leveraging the base Phi-2 model excels in the SQA-I and MM-Vet benchmarks. This superiority can likely be attributed to Phi-2’s extensive pretraining on large code datasets, which enhances its ability to generalize in science question-answering contexts and generate open-ended responses. While the performance differences among instruct-tuned models are modest, RLHF emerges as the most versatile performer across the eight benchmarks. Despite the close numerical results, qualitative analysis reveals a distinct divergence between instruct-tuned and base models. As depicted in Fig 3 (right), when tasked with describing an artwork image, the responses from models with base LM and SFT LM adopt a formal tone, employing terms like "religious elements" and "three-dimensional sculptures", reminiscent of textbook descriptions. Conversely, models finetuned with RLHF and Safe-RLHF LMs offer descriptions in a more human-like manner, akin to a museum commentary.

Take away. The benefit of employing an instruct-tuned LM for MSLMs over a base LM is not necessarily clear-cut, irrespective of the instruction tuning approach used, whether it be supervised fine-tuning, reinforcement learning from human feedback (RLHF), or safety-aware RLHF. This finding contrasts with prior research, which indicated that using an instruct-tune language model like Vicuna offers advantages over a base language model like LLaMA. However, **qualitative differences become apparent:** MSLMs developed using RLHF-based instruction tuning methods tend to be more verbose and produce responses that are more akin to human-like answers.

Visual Representations in MSLM

Pretrained Visual Representation. In multimodal models, a common approach is to employ a pretrained visual backbone that converts image features into latent embeddings. The quality of these pretrained embeddings is pivotal to the success of



User: Does this artwork belong to the type of study?

Base: The image features two young girls, possibly depicting a religious scene, and is described as a "statue of two girls." The presence of religious elements and the artistic style of the artwork suggest that it is part of an art history study.

SFT: Yes, this artwork belongs to the type of study called "statues." Statues are three-dimensional sculptures that represent human figures, often depicting people in various poses, expressions, and interactions. This type of artwork can be found in various settings, such as museums and galleries.

RLHF: The image features two young boys, possibly depicting a religious scene, and is described as a "statue of two boys." The artwork is displayed on a round surface, which is a common characteristic of statues and sculptures.

Safe-RLHF: Yes, this artwork belongs to the type of study called "statue study." In the image, two young children are depicted in a statue study, with one child looking at the other. The artist has captured the facial expressions and body language of the children.

Figure 3: **Base vs. Instruct-tuned LMs.** For MLLMs, we explore the differences using base LM and instruct-tuned LM. While the quantitative performance metrics appear similar, the qualitative results reveal differences (left). For example, when comparing the responses generated by MLLMs equipped with either a Base or SFT LM to those from models finetuned with RLHF or Safe-RLHF, the latter is observed to be more verbalized (right).

multimodal learning tasks. We evaluated the efficacy of four different pretrained visual backbones: 1) CLIP (Radford et al. 2021), trained on a large proprietary dataset by OpenAI, 2) SigLIP (Zhai et al. 2023), pretrained on the WebLI dataset, 3) ViT with IN21K, trained on the ImageNet21K (Ridnik et al. 2021) dataset using vision transformer (Dosovitskiy et al. 2021), and 4) DinoV2 (Oquab et al. 2023), pretrained on the LVD-142M dataset. For an equitable comparison, we standardized the use of ViT-Large with a 14-patch size and an input resolution of 336 pixels. As illustrated in Figure 4 (left), it becomes evident that training on extensive visual corpora—akin to CLIP’s methodology—confers an advantage in terms of enhanced visual understanding capabilities. *Take away.* The overarching conclusion from our experiments in this section resonates with insights gained from the training of large vision-language models. In particular, we find that the vision backbone, trained using vision-language contrastive loss (i.e., CLIP and SigLIP), outperforms other models. Interestingly, our experiments further reveal that SigLIP demonstrates a surprisingly significant performance enhancement over CLIP, despite being supervised and operating on a considerably smaller training dataset. This observation suggests that refining training methodologies for CLIP-based models could be more impactful for small Vision-Language Models (VLMs) than for their larger counterparts, indicating a potential shift in optimization strategies for enhancing model effectiveness across different scales.

Scaling Image Resolution. Image resolution plays a vital role among various design choices in neural network architecture, with past research (Lu et al. 2023) indicating that, for MLLMs, increasing image resolution is more effective than scaling other dimensions, such as model width and depth.

This assertion is supported by findings (Karamcheti et al. 2024) that enhancing image resolution uniformly improves model performance across all benchmarks. Naturally, one might assume that this principle should apply equally to MSLMs. However, is this assumption truly valid?

We present experiments using a popular visual backbone, CLIP, with varying image sizes. The experimental results are presented in Fig 4 (right). Contrary to the common belief that increasing image resolution is always beneficial to model generalization, our empirical results indicate that higher image resolution does not guarantee improved performance for MSLMs. For instance, a CLIP model with a 224px resolution outperformed its higher-resolution counterparts in specific benchmarks, while a 336px resolution CLIP model excelled in others. Despite these varied results, it’s acknowledged that increased resolution can benefit tasks requiring detailed visual understanding, such as OCR, where higher resolution inputs enable more accurate recognition of details like price tags in our qualitative example in the appendix.

Take away. **Scaling up image resolution is not a silver bullet.** Although higher image resolutions offer more detailed visual inputs, they do not invariably enhance the performance of MSLMs. It is essential to find the right equilibrium between resolution and computational efficiency, tailored to the model’s application context and the need for consistent understanding and output. Efforts might be better allocated to scaling other aspects of the visual backbone, such as its depth, rather than relying solely on increased image resolution.

Optimization

Optimization strategies play a pivotal role in shaping the generalizability of a model. Most MLLMs (Liu et al. 2023b,a;

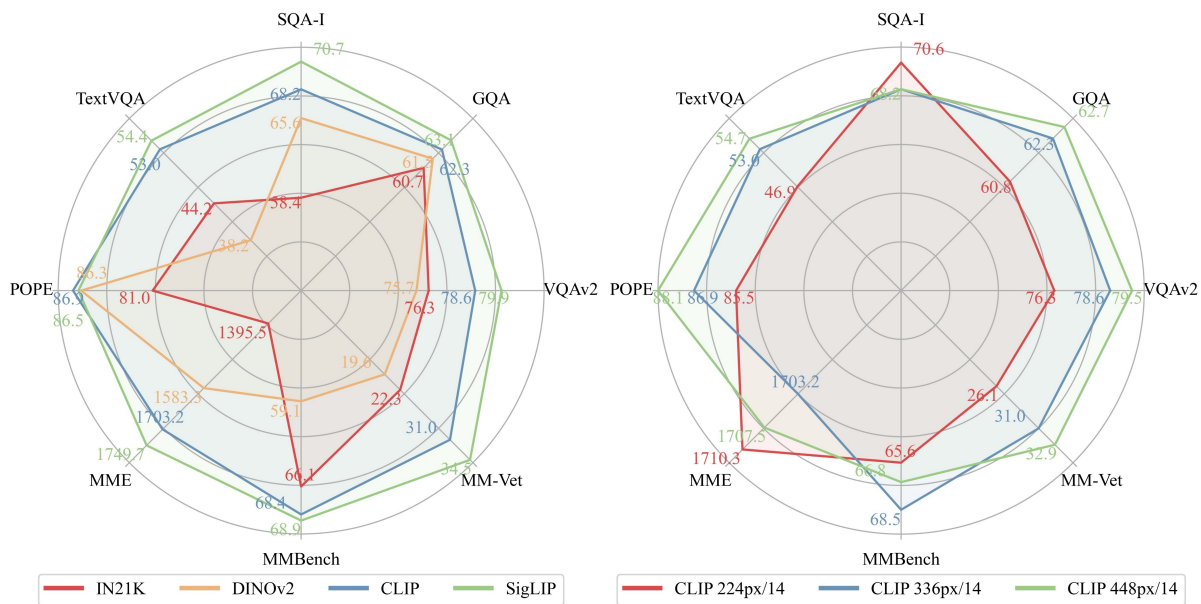


Figure 4: **Choosing a Pretrained Vision Representation & Scaling Image Resolution.** We evaluate various visual backbones, such as CLIP, SigLIP, DINOv2, and ViT-IN21K (left). We analyze model performance in relation to increasing image resolution (right). We provide examples from models employing visual backbones at different image resolutions in the appendix.

Zhu et al. 2024a; Ye et al. 2023) typically adopt a two-stage optimization pipeline. The first stage, or the pretraining stage, is dedicated to harmonizing vision and language features, establishing a solid foundation for multimodal understanding. The subsequent stage focuses on instructing the model to follow users’ instructions. This section delves into the training paradigms of MLLMs, exploring the necessity of finetuning the language model and visual backbones. Additionally, we evaluate the feasibility and effectiveness of parameter-efficient fine-tuning methods in this context.

Frozen vs. Finetuning. This section examines whether the visual representation backbone and the language model should be configured as trainable or frozen. We present and analyze four different configurations that were explored.

- **Frozen ViT+SLM:** The parameters of both the ViT and SLM are held constant, allowing the training to focus on aligning the two components via the projector.
- **Frozen SLM + Finetune ViT:** Here, we freeze the small language model’s parameters and solely fine-tune the Vision Transformer.
- **Finetune LM + Frozen ViT:** This strategy involves freezing the ViT while training both the projector and SLM, an approach exemplified by LLaVA-1.5 (Liu et al. 2023a).
- **Finetune ViT + SLM:** This approach activates all parameters within the MSLMs, permitting the model to simultaneously refine its visual processing capabilities and improve instruction-following proficiencies.

Our exploration into these strategies involves selectively activating or freezing the visual and language model during the instruction-tuning stage. Our findings suggest that freezing the language model can detrimentally affect the MSLMs’

performance across various benchmarks. Furthermore, we observe that fully engaging all parameters of the MSLMs can lead to improvements on certain benchmarks.

Take away. Contrary to previous research that suggested fine-tuning the vision backbone could lead to a significant decrease in performance across all benchmarks (Karamcheti et al. 2024), our study indicates that fine-tuning the vision encoder can actually enhance the performance of MSLMs across all evaluated benchmarks. We believe the underlying reason for this discrepancy is that SLMs possess fewer active neurons than LLMs. As a result, the visual representations in MSLMs must be more expressive than those in MLLMs to enable the language model to effectively comprehend image contents. Furthermore, our observations reveal that fine-tuning the language model is even more critical for achieving a successful MSLM implementation. We attribute this necessity to the fact that SLMs have less pre-trained data and fewer parameters, making the learning process to understand images with the assistance of visual backbones a pivotal challenge.

Full-Parameter Tuning vs. LoRA. The aforementioned section shows that fine-tuning the visual backbone and LMs is necessary for MSLMs. However, full-parameter fine-tuning entails a significant increase in training costs. We investigate the potential of parameter-efficient fine-tuning techniques as alternatives to conventional full-parameter tuning. We juxtapose the performance of models utilizing Low-Rank Adaptation (LoRA) against those subjected to full-parameter tuning. For the LoRA setup, we configure r to be 128 and α to be 256. The experimental outcomes suggest that LoRA attains performance on par with its full-parameter

Method	LM	Res.	PT	IT	VQAv2	GQA	SQA ^I	VQA ^T	MME-P	MME-C	MMB	MM-Vet	POPE
Multimodal Large Language Models													
BLIP-2	Vicuna (13B)	224	129M	-	65.0	41.0	61.0	42.5	1293.8	290.0	-	22.4	85.3
InstructBLIP	Vicuna (7B)	224	129M	1.2M	-	49.2	60.5	50.1	-	-	36	26.2	-
InstructBLIP	Vicuna (13B)	224	129M	1.2M	-	49.5	63.1	50.7	1212.8	291.8	-	25.6	78.9
Shikra	Vicuna (13B)	224	600K	5.5M	77.4*	-	-	-	-	-	58.8	-	-
IDEFICS-9B	LLaMA (7B)	224	353M	1M	50.9	38.4	-	25.9	-	-	48.2	-	-
IDEFICS-80B	LLaMA (65B)	224	353M	1M	60.0	45.2	-	30.9	-	-	54.5	-	-
Qwen-VL	Qwen (7B)	448	1.4B [†]	50M [†]	78.8*	59.3*	67.1	63.8*	-	-	38.2	-	-
Qwen-VL-Chat	Qwen (7B)	448	1.4B [†]	50M [†]	78.2*	57.5*	68.2	61.5*	1487.5	360.71	60.6	-	-
mPLUG-Owl2	LLaMA (7B)	448	400M	1.23M	79.4	56.1	68.7	58.2	1450.2	313.2	64.5	36.2	85.8
LLaVA-1.5	Vicuna (7B)	336	558K	665K	78.5*	62.0*	66.8	58.2	1510.7	316.1	64.3	30.5	85.9
LLaVA-1.5	Vicuna (13B)	336	558K	665K	80.0*	63.3*	71.6	61.3	1531.3	295.4	67.7	35.4	85.9
Multimodal Small Language Models													
MobileVLM-1.7B	M-LLaMA (1.4B)	336	558K	665K	-	56.1	57.3	41.5	1196.2	-	53.2	-	84.5
MobileVLM-3B	M-LLaMA (2.7B)	336	558K	665K	-	59.0	61.2	47.5	1288.9	-	59.6	-	84.9
MobileVLM-v2-1.7B	M-LLaMA (1.4B)	336	1.2M	2.4M	-	59.3	66.7	52.1	1302.8	-	57.7	-	84.3
MobileVLM-v2-3B	M-LLaMA (2.7B)	336	1.2M	2.4M	-	61.1	70.0	57.5	1440.5	-	63.2	-	84.7
MoE-LLaVA-3.6B	Phi-2 (2.7B)	384	558k	1.59M	79.9*	62.6*	70.3	57.0	1431.3	-	68.0	35.9	85.7
Bunny-3B	Phi-2 (2.7B)	384	2M	695K	79.8	62.5	70.9	-	1488.8	289.3	68.6	-	86.8
Mipha-1.6B	Phi-1.5 (1.3B)	384	558K	665K	77.5*	62.7*	58.3	45.6	1203.1	247.9	57.7	23.5	86.9
Mipha-2.4B	Gemma (2.0B)	384	558K	665K	79.5*	63.3*	65.3	52.4	1397.1	265.7	59.4	29.9	86.6
Mipha-3B	Phi-2 (2.7B)	384	558K	665K	81.3*	63.9*	70.9	56.6	1488.9	295.0	69.7	32.1	86.7

Table 2: Multi-modal evaluation on 9 benchmarks: VQAv2; GQA; SQA^I: ScienceQA-IMG; VQA^T: TextVQA; MME-P: MME perception; MME-C: MME cognition; MMB: MMBench; MM-Vet; POPE. *The training images/annotations of the datasets are observed during training. [†] Includes in-house data that is not publicly accessible. Column *Res.* is the image resolution of visual backbone. Columns *PT* and *IT* are the data sizes in the pretraining stage and the visual instruction tuning stage, respectively.

counterparts across all evaluated benchmarks, positioning it as a alternative to full-parameter fine-tuning.

Experiments

Experimental Results on Visual Question Answering Benchmarks. We evaluate the visual question answering abilities on VQAv2 (Goyal et al. 2017), GQA (Hudson and Manning 2019), VizWiz (Gurari et al. 2018), SQA^I (Lu et al. 2022) and VQA^T (Singh, Natarajan et al. 2019). As shown in Table 2, Mipha-3B achieves the highest performance in 2 out of the 5 benchmarks. Notably, when compared to Bunny-3B (He et al. 2024), which is trained on four times more data than Mipha-3B, our method demonstrates superior performance with improvements of 1.5% and 1.4% on VQAv2 and GQA, respectively. This underlines the effectiveness of our optimization strategies. Moreover, without the need for an expanded training dataset, Mipha-3B stands out on VQAv2 and GQA, outperforming LLaVA-1.5-13B (Liu et al. 2023a) on VQAv2 by 1.3% and on GQA by 0.6%. These findings underscore that Mipha-3B, despite having only 23% of the parameter count of LLaVA-1.5-13B, can still achieve commendable generalization in certain scenarios.

Experimental Results on Instruction-Following Benchmarks. Furthermore, we evaluate our proposed Mipha on more challenging benchmarks that are designed for evaluation of MLLMs, i.e., POPE (Li et al. 2023b), MME (Fu et al. 2023), MMBench (Liu et al. 2023d),

and MM-Vet (Yu et al. 2023). As shown in Table 2, our proposed model, Mipha-3B, exceeds LLaVA-1.5-13B in performance on MMBench by 2%, despite both models being trained on the same dataset. Additionally, Mipha-3B demonstrates stronger resistance to hallucinations, surpassing LLaVA-1.5-13B on POPE by 0.8%. In comparison with other MSLMs, Bunny-3B, MoE-LLaVA-3.6B (Lin et al. 2024), and MobileVLM-v2-3B (Chu et al. 2024), we are astonished to find that despite Mipha-3B utilizing the least amount of training data, it achieved the best performance in 4 out of 6 benchmarks. In particular, on MMBench, Mipha-3B outperformed Bunny-3B by 1.1% and MoE-LLaVA by 1.7%, which are combined with the same visual backbone SigLIP and LM Phi-2-2.7B respectively. These results validate the effectiveness of our optimization strategies.

Conclusion

In this paper, we delve deeply into three aspects of Multimodal Small Language Models (MSLMs): the language models, visual representations, and optimization strategy. Based on our thorough analysis, we identify several improvements that can enhance the performance of MSLMs. Integrating insights from these aspects, we have developed a novel suite of models called Mipha. Compared to other MSLMs, our top-performing model, Mipha-3B, secures the highest performance in 6 out of 9 benchmarks. We believe that Mipha offers a novel perspective on the training of strong MSLMs.

Acknowledgments

This work is supported by the Sci-Tech Innovation Initiative by the Science and Technology Commission of Shanghai Municipality (24ZR1419000), the National Science Foundation of China (12471501), and the Science and Technology Innovation Action Plan of Shanghai under Grant 22511105400.

References

- Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- 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. *Advances in Neural Information Processing Systems*, 35: 23716–23736.
- 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*.
- Biderman, S.; Schoelkopf, H.; Anthony, Q. G.; Bradley, H.; O’Brien, K.; Hallahan, E.; Khan, M. A.; Purohit, S.; Prashanth, U. S.; Raff, E.; et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2397–2430. PMLR.
- Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H. P. d. O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023).
- Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113.
- Chu, X.; et al. 2024. MobileVLM V2: Faster and Stronger Baseline for Vision Language Model. *arXiv preprint arXiv:2402.03766*.
- Cobbe, K.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Dai, W.; et al. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv preprint arXiv:2305.06500*.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Housley, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.
- Fu, C.; et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.
- Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6904–6913.
- Gunasekar, S.; Zhang, Y.; Aneja, J.; Mendes, C. C. T.; Del Giorno, A.; Gopi, S.; Javaheripi, M.; Kauffmann, P.; de Rosa, G.; Saarikivi, O.; et al. 2023. Textbooks Are All You Need. *arXiv preprint arXiv:2306.11644*.
- Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3608–3617.
- He, M.; Liu, Y.; Wu, B.; Yuan, J.; Wang, Y.; Huang, T.; and Zhao, B. 2024. Efficient Multimodal Learning from Data-centric Perspective. *arXiv preprint arXiv:2402.11530*.
- Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multi-task Language Understanding. In *International Conference on Learning Representations*.
- Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6700–6709.
- Javaheripi, M.; Bubeck, S.; Abdin, M.; Aneja, J.; Bubeck, S.; Mendes, C. C. T.; Chen, W.; Del Giorno, A.; Eldan, R.; Gopi, S.; et al. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog*.
- Karamcheti, S.; Nair, S.; Balakrishna, A.; Liang, P.; Kollar, T.; and Sadigh, D. 2024. Prismatic vlms: Investigating the design space of visually-conditioned language models. *arXiv preprint arXiv:2402.07865*.
- Li, J.; Zhu, Y.; Xu, Z.; Gu, J.; Zhu, M.; Liu, X.; Liu, N.; Peng, Y.; Feng, F.; and Tang, J. 2024. Mmro: Are multimodal llms eligible as the brain for in-home robotics? *arXiv preprint arXiv:2406.19693*.
- Li, Y.; Bubeck, S.; Eldan, R.; Del Giorno, A.; Gunasekar, S.; and Lee, Y. T. 2023a. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.
- Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.
- Lin, B.; et al. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.
- Lin, T.-Y.; Maire, M.; et al. 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, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual Instruction Tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Liu, X.; Zhu, Y.; Gu, J.; Lan, Y.; Yang, C.; and Qiao, Y. 2025. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, 386–403. Springer.
- Liu, X.; Zhu, Y.; Lan, Y.; Yang, C.; and Qiao, Y. 2023c. Query-relevant images jailbreak large multi-modal models. *arXiv preprint arXiv:2311.17600*.
- Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2023d. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.
- Lu, P.; et al. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521.
- Lu, Y.; Li, C.; Liu, H.; Yang, J.; Gao, J.; and Shen, Y. 2023. An empirical study of scaling instruct-tuned large multimodal models. *arXiv preprint arXiv:2309.09958*.
- Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.
- 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.
- Ridnik, T.; Ben-Baruch, E.; Noy, A.; and Zelnik-Manor, L. 2021. ImageNet-21K Pretraining for the Masses. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.
- Singh, A.; Natarajan, V.; et al. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8317–8326.
- Srivastava, A.; et al. 2022. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.
- Stability. 2024. StableLM: Stability AI Language Models.
- Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; XiXuan, S.; et al. 2023. CogVLM: Visual Expert for Large Language Models. *arXiv preprint*.
- Wen, J.; Zhu, M.; Zhu, Y.; Tang, Z.; Li, J.; Zhou, Z.; Li, C.; Liu, X.; Peng, Y.; Shen, C.; et al. 2024a. Diffusion-VLA: Scaling Robot Foundation Models via Unified Diffusion and Autoregression. *arXiv preprint arXiv:2412.03293*.
- Wen, J.; Zhu, Y.; Li, J.; Zhu, M.; Wu, K.; Xu, Z.; Liu, N.; Cheng, R.; Shen, C.; Peng, Y.; et al. 2024b. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *arXiv preprint arXiv:2409.12514*.
- Wen, J.; Zhu, Y.; Zhu, M.; Li, J.; Xu, Z.; Che, Z.; Shen, C.; Peng, Y.; Liu, D.; Feng, F.; and Tang, J. 2024c. Object-Centric Instruction Augmentation for Robotic Manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 4318–4325.
- Ye, Q.; et al. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*.
- Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; and Wang, L. 2023. Mm-vet: Evaluating large multi-modal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.
- Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*.
- Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2024a. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *The Twelfth International Conference on Learning Representations*.
- Zhu, M.; Zhu, Y.; Li, J.; Wen, J.; Xu, Z.; Che, Z.; Shen, C.; Peng, Y.; Liu, D.; Feng, F.; and Tang, J. 2024b. Language-Conditioned Robotic Manipulation with Fast and Slow Thinking. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 4333–4339.
- Zhu, M.; Zhu, Y.; Li, J.; Wen, J.; Xu, Z.; Liu, N.; Cheng, R.; Shen, C.; Peng, Y.; Feng, F.; et al. 2024c. Scaling diffusion policy in transformer to 1 billion parameters for robotic manipulation. *arXiv preprint arXiv:2409.14411*.
- Zhu, Y.; Ou, Z.; Mou, X.; and Tang, J. 2024d. Retrieval-Augmented Embodied Agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17985–17995.
- Zhu, Y.; Zhu, M.; Liu, N.; Ou, Z.; Mou, X.; and Tang, J. 2024e. LLaVA-Phi: Efficient Multi-Modal Assistant with Small Language Model. *arXiv preprint arXiv:2401.02330*.