

Cobra: Extending Mamba to Multi-Modal Large Language Model for Efficient Inference

Han Zhao^{1 2*}, Min Zhang^{1*}, Wei Zhao², Pengxiang Ding², Siteng Huang³, Donglin Wang^{2†}

¹Zhejiang University

²Westlake University

³DAMO Academy, Alibaba Group

zhaohan34@westlake.edu.cn, zhangmin@westlake.edu.cn, zhaowei@westlake.edu.cn,
dingpengxiang@westlake.edu.cn, siteng.huang@gmail.com, wangdonglin@westlake.edu.cn

Abstract

In recent years, applying multi-modal large language models (MLLMs) in various fields has achieved remarkable success. However, as the foundation model for many downstream tasks, MLLMs comprise the well-known Transformer network, which has a less efficient quadratic computation complexity. In this paper, we introduce Cobra, a multi-modal large-scale language model built upon a state-space model, which has demonstrated significant potential in efficiently handling long sequences with fast inference and linear scalability concerning sequence length. Specifically, Cobra involves replacing Transformer-based backbone models (*e.g.*, LLaMA or Phi) with pre-trained Mamba language models. We then empirically explore effective strategies for aligning visual and textual modalities and integrating various pre-trained Mamba model variants with visual encoders. Experiments across various multi-modal benchmarks demonstrate that: (i) Cobra performs $3\times \sim 4\times$ faster than the most computationally efficient state-of-the-art methods, *e.g.*, LLaVA-Phi and MobileVLM v2. Additionally, its performance is significantly enhanced thanks to the implementation of linear sequential modeling. (ii) Cobra fine-tunes a small parameter ($\sim 48\%$ of model parameters), leading to a significant improvement in overall performance compared to LLaVA.

Code — <https://github.com/h-zhao1997/cobra>

1 Introduction

Multi-modal large language models (MLLMs) have recently achieved impressive results across a variety of downstream tasks, including multi-modal content generation (Lu et al. 2022; Wu et al. 2024), vision-based question answering (OpenAI 2023; Gao et al. 2023; Liu et al. 2023b,c), and embodied intelligence (Brohan et al. 2023; Kim et al. 2024; Ding et al. 2024). By effectively aligning pre-trained large language models with visual modalities, MLLMs have demonstrated a strong ability to comprehend and navigate complex visual-language contexts. These advancements not only highlight the versatility of MLLMs but also pave the

way for further research into more nuanced and sophisticated applications. For example, as MLLMs continue to evolve, there is potential for significant improvements in areas such as real-time interaction in dynamic environments, cross-modal retrieval tasks, and the seamless integration of language and visual processing in everyday technology.

MLLMs typically rely on the well-known Transformer network as a foundational model for many downstream tasks. However, due to their quadratic computational complexity, Transformer networks are often less efficient, making it challenging to meet the demands of application scenarios that require high real-time performance and are suitable for edge deployment. As shown in Figure 1 (d), the MoE-LLaVA (Lin et al. 2024) model can process only 20.33 tokens generated per second, indicating a low processing efficiency. Despite these challenges, there remains a significant demand for MLLMs in such areas. Therefore, **the ability to deploy MLLMs that support fast inference with low resource utilization is particularly crucial.**

Traditional approaches have primarily focused on improving the efficiency of MLLMs by reducing model capacity or compressing the length of the visual context while generally maintaining the Transformer architecture within the language model (Zhu et al. 2024; Zhou et al. 2024; Zhang et al. 2024; Chu et al. 2023, 2024; Lin et al. 2024). For example, LLaVA-Phi (Zhu et al. 2024) builds a multi-modal base model using a small-scale Phi-2 as the core language model. MobileVLM (Chu et al. 2023, 2024) utilizes MobileLLaMA as its base model, training a series of smaller-scale language models based on the LLaMA architecture. These methods aim to enhance the inference speed of MLLMs by reducing the size of the language models. Although this approach improves efficiency, it often comes at the expense of significantly reduced model performance. In Figure 1 (b), MobileVLM v2-3B (Chu et al. 2024) has the worst performance on all MLLM benchmarks compared to the other models.

In this paper, our primary goal is to enhance the inference speed of multi-modal large language models (MLLMs) while ensuring their performance remains uncompromising. To achieve this, we propose Cobra, which integrates the Mamba large language model and utilizes the linear scalability of state-space modeling (SSM). This approach effectively addresses the quadratic computational complex-

*These authors contributed equally.

†Corresponding author.

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

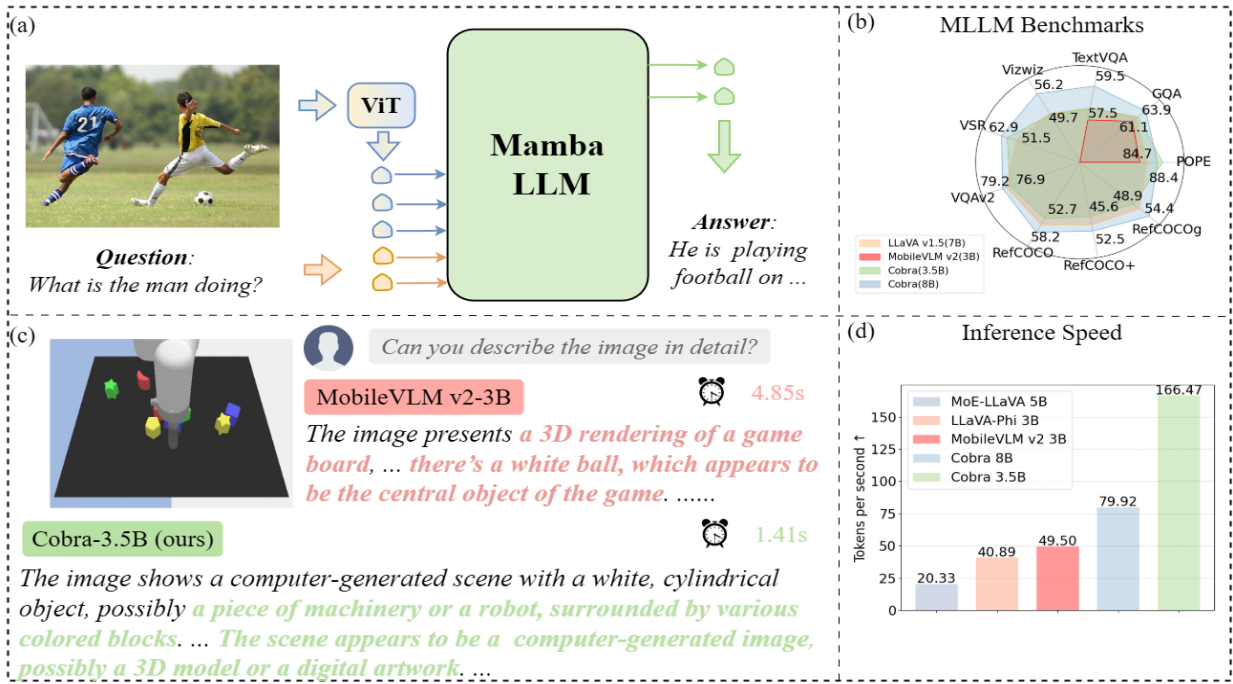


Figure 1: **Overview of Cobra.** (a) Our innovative integration of a vision encoder with the efficient Mamba language model significantly enhances the reasoning efficiency of MLLMs. (b) Cobra demonstrates competitive performance on general MLLM benchmark tests. (c) Cobra generates accurate textual descriptions (e.g., green text indicates a correct answer), outperforming current state-of-the-art MLLMs that produce inaccurate answers (shown in red text) while maintaining rapid reasoning speeds. (d) A comparison of our proposed Cobra and the baseline in terms of the number of tokens generated per second.

ity inherent in traditional Transformer architectures. Specifically, Cobra consists of three key components: a vision encoder that concatenates DINOv2 (Oquab et al. 2024) and SigLIP (Zhai et al. 2023) features, a projector that maps visual features to the language embedding space and the Mamba LLM backbone, as shown in Figure 1 (a) and Figure 2. In Figure 1 (c), Cobra performs $3\times \sim 4\times$ faster than MobileVLM v2-3B. Interestingly, Cobra can generate more accurate responses, as highlighted by the text in green, effectively mitigating the hallucination problem commonly seen in MLLMs. Even compared to the much larger LLaVA v1.5 model with 7 billion parameters, Cobra still performs comparably on several specific benchmarks with about 48% of the parameters. Our main contributions are summarised:

- We investigate that existing MLLMs typically rely on Transformer networks, exhibiting a quadratic computational complexity. To address this inefficiency, we present Cobra, a novel MLLM with linear computation.
- Our research investigates various multi-modal fusion strategies to enhance the integration of visual and language within the Mamba LLM. Extensive experiments evaluate the effectiveness of different fusion approaches.
- Extensive experiments are conducted to evaluate the performance of Cobra in comparison to concurrent studies aimed at improving the computational efficiency of foundational MLLMs. Notably, Cobra-3.5B even achieves comparable performance to LLaVA with fewer parameters, underscoring its efficiency. Cobra-8B surpasses the

LLaVA v1.5 model of similar size on all tested benchmarks, achieving an average accuracy improvement of approximately 6%. It also remains faster in inference compared to MobileVLM v2-3B.

2 Related Works

2.1 Multi-Modal Large Language Models

Building on the success of large language models (LLMs), numerous extensions have been developed to apply LLMs to multi-modal tasks, integrating information from multiple sources such as text, images, and audio to enable comprehensive understanding and reasoning across different modalities (Chu et al. 2023; Liu et al. 2023b; Taori et al. 2023; Bai et al. 2023; Alayrac et al. 2022; Awadalla et al. 2023; Liu et al. 2023c; Chen et al. 2023b). These models leverage vast amounts of data and intricate architectures to achieve state-of-the-art performance in tasks such as image captioning (Ke et al. 2019), visual question answering (Antol et al. 2015) and cross-modal retrieval (Hendriksen et al. 2023). Recent advances have harnessed the formidable reasoning power of LLMs such as LLaMA (Touvron et al. 2023) and Vicuna (Chiang et al. 2023). However, a notable commonality among existing MLLMs is their reliance on the Transformer backbone to model dependencies among sequential tokens. Despite the Transformer network’s exceptional capability in capturing relationships within data, its quadratic computational complexity presents a significant drawback, particularly when dealing with large-scale language models.

To mitigate this problem, several studies have been proposed to present more compact and efficient MLLMs (Zhu et al. 2024; Zhou et al. 2024; Zhang et al. 2024; Chu et al. 2023, 2024). For example, LLaVA-Phi (Zhu et al. 2024) builds a multi-modal foundation model taking the small-scale Phi-2 as the LLM. MobileVLM (Chu et al. 2023, 2024) introduces MobileLLaMA as the base model which trains a family of small-scaled LLM based on LLaMA architecture. However, these methods achieve effective MLLMs by using smaller-scale LLMs that significantly reduce the performance while increasing the speed of inference.

2.2 State Space Models

State space models (SSMs) have demonstrated highly promising performance across various tasks, including long-range sequence modeling (Smith, Warrington, and Linderman 2023; Hasani et al. 2022), image generation (Yan, Gu, and Rush 2023; Bellagente et al. 2024) and reinforcement learning (Bar-David et al. 2023; Lu et al. 2023). One of the key advantages of SSMs is their flexibility, as they can be formulated as recurrent neural networks (RNNs) for efficient inference or as models capable of parallel processing entire input sequences, enabling more efficient training.

Recently, a new selective SSM structure called Mamba (Gu and Dao 2023) has been introduced, which is regarded as a strong competitor to the Transformer architecture. Compared to LLMs of similar capacity, Mamba-based language models (Dao and Gu 2024; Waleffe et al. 2024) demonstrate competitive performance with a distinct advantage: their inference speeds scale linearly with sequence length while maintaining constant memory usage. This efficiency allows Mamba to handle long contexts and perform inference more effectively. In contrast, Transformer-based models face challenges such as increasing GPU memory consumption and computation time that grows quadratically with sequence length (Katharopoulos et al. 2020). In this paper, we conduct an in-depth exploration of extending Mamba-based LLMs into practical and efficient MLLMs. Through extensive experiments, we examine the distinctive characteristics of Mamba MLLMs and develop efficient training strategies to significantly enhance their performance.

3 Methodology

This section introduces the preliminary concepts of state space models (Section 3.1). Then we describe the details of our proposed Cobra (Section 3.2), which mainly includes the vision encoder, the projector, and the Mamba LLM.

3.1 Preliminaries

Traditional state space models (SSMs) (Gu, Goel, and Ré 2022; Smith, Warrington, and Linderman 2023) are characterized by the parameters (Δ, A, B, C) . Given a continuous-time scalar input signal $x(t)$, the SSM can be described by the following ordinary differential equation:

$$h'(t) = Ah(t) + Bx(t), \quad (1a)$$

$$y(t) = Ch(t), \quad (1b)$$

where the parameters of the SSM, $A \in \mathbb{R}^{N \times N}$, $B \in \mathbb{R}^{N \times 1}$ and $C \in \mathbb{R}^{1 \times N}$, represent constant matrices. The variables h , x , and y are continuous-time variables concerning time t , representing the hidden state, input, and output.

In practice, the SSMs operate in a discretized form to handle input sequences, and we use a mixer layer that constructs an SSM for each input channel independently. Δ is a time-scale parameter that helps transform A and B into discrete-time parameters \bar{A} and \bar{B} , respectively. The discretization rule for A and B with the zero-order hold is as follows:

$$\bar{A} = \exp(\Delta A), \quad (2a)$$

$$\bar{B} = (\Delta A)^{-1}(\exp(\Delta A) - I) \cdot \Delta B. \quad (2b)$$

Thus, the structured SSM can be summarized as the following recurrence form in Equation(3a) and (3b):

$$h_k = \bar{A}h_{k-1} + \bar{B}x_k, \quad (3a)$$

$$y_k = C\bar{h}_k. \quad (3b)$$

The model can also be written as the convolution form (4a), (4b) to process the sequence in parallel:

$$\bar{K} = (C\bar{B}, C\bar{A}\bar{B}, \dots, C\bar{A}^k\bar{B}, \dots) \quad (4a)$$

$$y = x * \bar{K} \quad (4b)$$

Based on the structured SSM, the selective SSM (Gu and Dao 2023) is further introduced to endow the model with the ability to selectively propagate or forget information according to the sequential input tokens. Specifically, the selective SSM achieves these functions by making the parameters (\bar{A}, \bar{B}, C) depend on the input x , which significantly enhances the model’s expressive capacity. Gu et al. (Gu and Dao 2023) proposed a hardware-aware algorithm called selective scan to allow efficient implementation of the model.

3.2 Cobra Model

To accomplish the purpose of building a multi-modal large language model (MLLM) that is capable of receiving visual information, we introduce Cobra as illustrated in Figure 2. Cobra consists of three key components: a vision encoder, a projector, and a Mamba LLM backbone. We present the implementation details for each component below.

- Vision encoder: We fuse DINOv2 (Oquab et al. 2024) and SigLIP (Zhai et al. 2023) as our vision backbone. The intuition is that combining the visual representations, which capture low-level spatial properties from DINOv2 and the semantic properties provided by SigLIP further improves the performance on downstream tasks (Tong et al. 2024; Karamcheti et al. 2024). Considering an image $X_v \in \mathbb{R}^{C \times H \times W}$ as input, the vision encoder splits the image into $N_v = HW/P^2$ same-size patches, where P^2 is the patch size. Both two vision encoders take the patchified image as an input token sequence and extract the channel-wise concatenation of the output of two encoders as the compact visual representations $R_v \in \mathbb{R}^{N_v \times (D_{\text{DINOv2}} + D_{\text{SigLIP}})}$:

$$R_v = [\varphi_{\text{DINOv2}}(X_v); \varphi_{\text{SigLIP}}(X_v)], \quad (5)$$

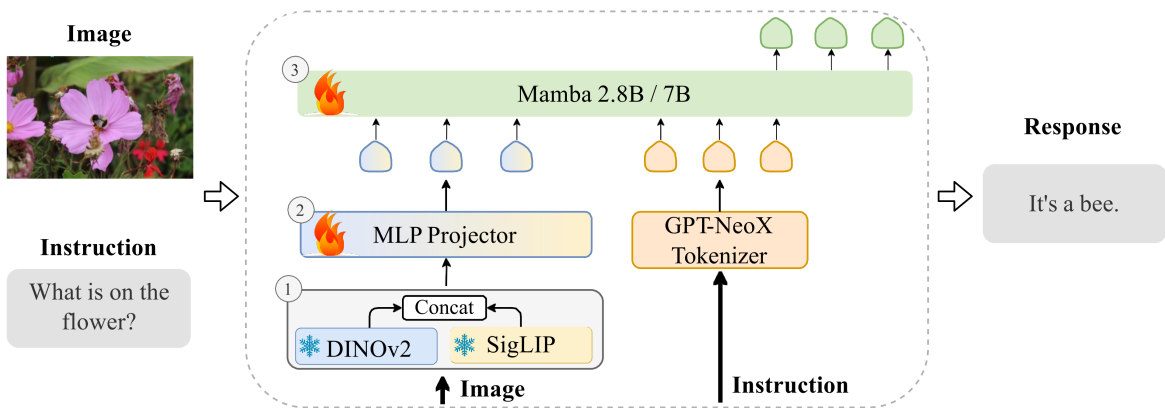


Figure 2: **Cobra model architecture.** Given an image observation and a language instruction, the model generates the corresponding answer. The architecture consists of three key components: ① a vision encoder that concatenates DINOv2 (Oquab et al. 2024) and SigLIP (Zhai et al. 2023) features, ② a projector that maps visual features to the language embedding space and ③ the Mamba LLM backbone, a Mamba 2.8 or 7B-parameter large language model (Gu and Dao 2023; Mercat et al. 2024).

for a subsequent task-specific head, where D_v denotes the dimension of the above-produced tokens.

- **Projector:** The projector is a simple learnable module that aligns the feature of vision and text by transforming the dimension of the original visual representation to the dimension of the tokens in the Mamba language model:

$$H_v = \phi(R_v). \quad (6)$$

We introduce two implementations of the different projectors in Cobra to map visual tokens into the same latent space as the language tokens. The multiple-layer perceptron (MLP) can be used to merge information from different modalities. In addition, the lightweight down-sample projector suggested by (Chu et al. 2024) is also tested to achieve a greater reduction in computation cost.

- **Mamba backbone:** The Mamba backbone is a stack of multiple identical basic blocks with the short convolution, SSM module, the residual connection, and RMSNorm (Zhang and Sennrich 2019) for each block. The model receives the concatenation of visual embeddings transformed from the projection layer and text embeddings, denoted as $H \in \mathbb{R}^{L_{in} \times D}$, and transforms this sequence into target token sequence $Y = \{y_i\}_{i=1}^L$ in an auto-regressive manner:

$$p(Y|H_v, H_q) = \prod_{i=1}^L p(y_i|H_v, H_q, y_{<i}). \quad (7)$$

Lastly, the tokens will be detokenized to the response answer in natural language.

3.3 Training Recipe

Recent research (Karamcheti et al. 2024) suggests that the pre-alignment phase may be unnecessary in the LLaVA-based training paradigm (Liu et al. 2023b; Chu et al. 2024) (*i.e.*, training only the pre-alignment phase of the projection layer and fine-tuning the large language model (LLM) for one epoch each). It has been observed that even after fine-tuning, the model remains underfitted. In light of this, we

Vision Encoder	DINOv2 + SigLIP ViT-SO
LLM init.	Mamba-2.8b-Zephyr / Mamba-7B
Projector init.	Random
Image resolution	384×384
Image token num.	729
Global batch size	128
Training steps	19K
Optimizer	AdamW
LR schedule	Cosine decay
Learning Rate	$2e-5$
Weight decay	0.1
Warm-up ratio	0.03
Number of epochs	2

Table 1: The configuration of models and hyperparameters.

chose to eliminate the pre-alignment stage and instead directly fine-tune the entire LLM backbone and the projector. This fine-tuning process spans two epochs, with random sampling conducted on a combined dataset comprising:

1. The mixed dataset used in LLaVA v1.5, which contains a total of 655K visual multi-turn conversations including academic VQA (Goyal et al. 2017; Hudson and Manning 2019; Krishna et al. 2016; Singh et al. 2019) samples, as well as visual instruction tuning data in LLaVA-Instruct (Liu et al. 2023c) and pure text instruction tuning data in ShareGPT (ShareGPT 2023).

2. LVIS-Instruct-4V (Wang et al. 2023), which contains 220K images with visually aligned and context-aware instructions generated by GPT-4V.

3. LRV-Instruct (Liu et al. 2023a), a 400K visual instruction dataset that covers 16 vision-and-language tasks aiming on mitigating hallucination.

Overall, the entire dataset contains approximately 1.2 million images, corresponding multi-turn dialogue data, and pure text dialogue data.

Model	LLM Backbone	Res.	VQA-v2	GQA	VizWiz	TextVQA	VSR	POPE
<i>Large Scale MLLMs</i>								
OpenFlamingo (Awadalla et al. 2023)	MPT-7B	336	52.7	-	27.5	33.6	-	-
BLIP-2 (Li et al. 2023a)	Vicuna-13B	224	-	41.0	19.6	42.5	50.9	-
MiniGPT-4 (Zhu et al. 2023)	Vicuna-7B	224	32.2	-	-	-	-	-
InstructBLIP (Dai et al. 2023)	Vicuna-7B	224	-	49.2	34.5	50.1	<u>54.3</u>	-
InstructBLIP (Dai et al. 2023)	Vicuna-13B	224	-	49.5	33.4	50.7	52.1	-
Shikra (Chen et al. 2023a)	Vicuna-13B	224	77.4	-	-	-	-	-
IDEFICS (Laurençon et al. 2023)	LLaMA-7B	224	50.9	-	35.5	25.9	-	-
IDEFICS (Laurençon et al. 2023)	LLaMA-75B	224	60.0	-	36.0	30.9	-	-
Qwen-VL (Bai et al. 2023)	Qwen-7B	448	78.2	59.3	35.2	63.8	-	-
LLaVA v1.5 (Liu et al. 2023b)	Vicuna-7B	336	<u>78.5</u>	<u>62.0</u>	<u>50.0</u>	<u>58.2</u>	51.5	<u>85.9</u>
Cobra-8B (ours)	Mamba-7B	384	79.2	63.9	56.2	59.5	62.9	87.6
<i>Small Scale MLLMs</i>								
MoE-LLaVA (Lin et al. 2024)	StableLM-1.6B	336	76.7	60.3	36.2	50.1	-	<u>85.7</u>
MoE-LLaVA (Lin et al. 2024)	Phi2-2.7B	384	79.9	62.6	<u>43.7</u>	57.0	-	<u>85.7</u>
LLaVA-Phi (Zhu et al. 2024)	Phi2-2.7B	336	71.4	-	35.9	48.6	-	85.0
MobileVLM v2 (Chu et al. 2024)	MobileLLaMA-2.7B	336	-	61.1	-	<u>57.5</u>	-	84.7
Cobra-3.5B (ours)	Mamba-2.8B	384	<u>77.8</u>	<u>62.3</u>	49.7	58.2	58.4	88.4

Table 2: Experiments of four open-ended benchmarks (blue) and two closed-set benchmarks (red). Res. represents the image resolution used for the vision encoder input. The best performance is highlighted in bold and the second-best result is underlined.

4 Experiments

In this section, we conduct extensive experiments to evaluate the performance of our proposed Cobra method, aiming to answer the following questions: **RQ1**: How does the performance of our proposed Cobra method compare with state-of-the-art MLLMs? (Section 4.2) **RQ2**: How does the inference speed of Cobra compare to three transformer-based baselines? (Section 4.3) **RQ3**: How effective is the proposed Cobra in different settings (or ablation study)? (Section 4.4)

4.1 Experimental Setup

Datasets. We conduct our experiments on a diverse set of nine benchmarks, including (1) four open-ended visual question answering (VQA), *i.e.*, VQA-v2 (Goyal et al. 2017), GQA (Hudson and Manning 2019), VizWiz (Gurari et al. 2018) and TextVQA (Singh et al. 2019). (2) two closed-set visual question answering (VQA), *i.e.*, VSR (Liu, Emerson, and Collier 2023) and POPE (Li et al. 2023b). (3) three visual grounding, *i.e.*, RefCOCO, RefCOCO+ and RefCOCOg (Kazemzadeh et al. 2014; Yu et al. 2016).

VQA-v2 (Goyal et al. 2017) evaluates models’ general ability to understand and reason about images and questions. GQA (Hudson and Manning 2019) assesses spatial understanding and multi-step inference in real-world images. VizWiz (Gurari et al. 2018) is similar to VQA-v2 but includes a series of unanswerable questions. TextVQA (Singh et al. 2019) focuses on reasoning from text in images. VSR (Liu, Emerson, and Collier 2023) is composed of demanding True/False questions that probe individual spatial relationships within various scenes, which is challenging to MLLMs. POPE (Li et al. 2023b) is comprised of specific Yes/No questions designed to evaluate MLLMs’ tendency to generate hallucinations. RefCOCO focuses on short descriptions with spatial anchors, RefCOCO+ relies on appearance-

based descriptions, and RefCOCOg emphasizes long and rich descriptions (Kazemzadeh et al. 2014; Yu et al. 2016).

Baseline methods. We compare Cobra to a large number of algorithms that span different sizes, including (1) large-scale MLLMs: OpenFlamingo (Awadalla et al. 2023), BLIP-2 (Li et al. 2023a), MiniGPT-4 (Zhu et al. 2023), InstructBLIP (Dai et al. 2023), Shikra (Chen et al. 2023a), IDEFICS (Laurençon et al. 2023), Qwen-VL (Bai et al. 2023) and LLaVA v1.5 (Liu et al. 2023b). (2) small-scale MLLMs: MoE-LLaVA (Lin et al. 2024), LLaVA-Phi (Zhu et al. 2024) and MobileVLM v2 (Chu et al. 2024).

Implementation details. Our training process includes multi-modal instruction tuning, during which we fine-tune both the multi-modal projector and the Mamba LLM. The model is trained using 8 NVIDIA A100 80GB GPUs. We have selected various open-source model weights, including Mamba with 2.8 billion and 7 billion parameters, to serve as the LLM backbone for our model. The model configurations and hyperparameters are detailed in Table 1, with additional information provided in the supplementary material.

Prompt order. In our prompt template design, we were surprised to discover that the word order in the templates significantly impacts the model’s performance, particularly in TextVQA. For example, Cobra, which follows LLaVA and InstructBLIP evaluations, uses input tokens parsed by the OCR system as prompts—formatted as “**Question**\n**Reference OCR token: ...**”. We found that this specific prompt structure reduced performance substantially, from 47.9% to 43.0%, compared to not using any prompts at all. Through extensive experimental exploration, we addressed this issue by adjusting the prompt order to “**Reference OCR token: ...**\n**Question**”, which improved performance. This sensitivity to prompt order may be due to an inductive bias in the RNN model. We hope that our findings will encourage

Model	LLM Backbone	Total Params	Visual Tokens	$Eval_{avg}$ (tokens/s)	Total (s)
MoE-LLaVA	Phi-2-2.7B	5.3B (3.6B Activated)	576	20.33	12.59
LLaVA-Phi	Phi-2-2.7B	3.1B	576	40.89	6.26
MobileVLM v2	MobileLLaMA-2.7B	3.1B	144	49.50	5.17
Cobra-3.5B (ours)	Mamba-2.8B	3.5B	729	166.47	1.54
Cobra-LDPv2-3.5B (ours)	Mamba-2.8B	3.5B	196	166.85	1.53
Cobra-8B (ours)	Mamba-7B	7.8B	729	79.92	3.20

Table 3: Latency comparison of small-scale MLLMs with ~ 3 billion parameters.

Model	RefCOCO	RefCOCO+	RefCOCog	Avg.
LLaVA v1.5	55.1	49.5	50.9	51.8
Cobra-3.5B	52.7	45.6	46.9	48.4
Cobra-8B	58.2	52.5	54.4	55.0

Table 4: Experiments of three visual grounding benchmarks. Avg. represents the average accuracy of the model on three benchmarks. The best performance is highlighted in bold.

further research in the community on this problem.

4.2 Overall Performance

In Table 1, we report the overall performance of Cobra and fourteen baselines under large-scale MLLMs and small-scale MLLMs on six datasets. According to Table 1, we have the following findings: (1) **In the large scale MLLMs setting with more than 7 billion parameters.** our proposed Cobra-8B achieves the best performance on all evaluated benchmarks. (2) **In the small scale MLLMs setting with around 3 billion (total or activated) parameters.** Cobra-3.5B achieved the best performance on all benchmarks except VQA-v2 and GQA, where it was only surpassed by MoE-LLaVA, a multi-modal mixture-of-experts language model expanded and fine-tuned from phi-2-2.7B. Our model lags behind by 1%-2% in accuracy on these two metrics, while our inference speed is over 8 times faster than that of the model as shown in Section 4.3.

It is noteworthy that Cobra-3.5B, with only 48% of the total parameters of LLaVA v1.5-7B, achieves comparable results on open-ended VQA benchmarks and shows significant improvements in the challenging closed-set prediction tasks of VSR and POPE. On these two benchmarks, there are performance improvements of 6.9% and 2.5% respectively.

As shown in Table 4, we also evaluated the localization capabilities of our two models alongside LLaVA v1.5-7B. The results indicate that Cobra-3.5B has accuracy rates that are 3%-4% lower than LLaVA v1.5-7B across all three benchmarks. In contrast, Cobra-8B exhibits the highest accuracy among the three models, outperforming the others by over 3% in accuracy on all benchmarks. Given that the training schemes for Cobra were identical, these results demonstrate that the grounding ability of the model is significantly influenced by the performance of the language model itself.

4.3 Inference Speed

We evaluated the generation speed of our model compared to three transformer-based baseline models of similar activated

parameter scales with different architectures: MoE-LLaVA, LLaVA-Phi, and MobileVLM v2.

In the evaluation, all models received the same example image. We used the same question “Describe the image specifically” as the textual prompt and set the number of output tokens to 256 for all models. The total time T_{total} from the image encoding to finished generating the complete answer is recorded and we calculated the average number of tokens generated per second by $Eval_{avg} = 256/T_{total}$.

All the evaluations were done on the hardware with a single Nvidia A100 PCIe 80GB GPU. The results from Table 3 show that our model has a significant advantage in inference speed compared to transformer-based models. Compared to MobileVLM v2, which has undergone several lightweight optimizations, Cobra only took about 30% of the time to complete inference when the number of visual tokens processed significantly increased.

We also evaluated the results of Cobra-LDPv2, a variant of our model that replaces the projector with an LDPv2 block, which reduces the number of visual tokens per image to 196. The results showed no significant speed improvement in our evaluation method. Due to the nature of parallel RNN models, the number of prompt tokens only affects the speed of the model’s first parallel forward process. Given that LDP significantly compresses visual information through pooling, it can impact the performance of MLLM to some extent (see our ablation studies for performance comparison). We believe that for the structure of Cobra or other RNN-based MLLMs, adopting such a lightweight design on the projector may be unnecessary.

4.4 Ablation Studies

We conduct ablation studies to verify the network design of Cobra, mainly involving the choice of projectors, vision encoders, language models, and training strategies.

Vision encoders. Recent works discover that despite CLIP-like language-image models may offer rich semantics, it has the potential to lose the detailed information for images themselves. Therefore, we adopt DINOv2 as a supplementary encoder and concatenate visual representations from two encoders for the subsequent LLM. As shown in Table 3, the fusion of DINOv2 and SigLIP features leads to better performance compared with SigLIP-only on all the benchmarks except TextVQA. Especially, we found the fused architecture significantly improves the accuracy by 5%-6% on VSR and localization benchmarks. This result implies that there is a meaningful principle when selecting the vision encoder for downstream tasks.

Model	VQA-v2	GQA	VizWiz	TextVQA	VSR	POPE	RefCOCO	RefCOCO+	RefCOCog
Cobra-3.5B	77.8	62.3	49.7	58.2	58.4	88.4	52.7	45.6	48.9
w/ SigLIP	77.5 (0.3 ↓)	61.8 (0.5 ↓)	48.3 (1.4 ↓)	58.8 (0.6 ↑)	53.2 (5.2 ↓)	88.2 (0.2 ↓)	46.7 (6.0 ↓)	40.1 (5.5 ↓)	43.8 (5.1 ↓)
w/ LDPv2	76.2 (1.6 ↓)	61.9 (0.4 ↓)	50.2 (0.5 ↑)	54.7 (3.5 ↓)	56.1 (2.3 ↓)	87.7 (0.7 ↓)	50.3 (2.4 ↓)	42.9 (2.7 ↓)	46.9 (2.0 ↓)
w/ Base	77.8 (0.0 ↑)	62.7 (0.4 ↑)	47.2 (2.5 ↓)	57.9 (0.3 ↓)	54.4 (4.0 ↓)	89.0 (0.6 ↑)	52.2 (0.5 ↓)	45.6 (0.0 ↑)	48.6 (0.3 ↓)
w/ 1 Ep FT	76.5 (1.3 ↓)	60.9 (1.4 ↓)	48.5 (1.2 ↓)	57.5 (0.7 ↓)	53.8 (4.6 ↓)	88.1 (0.3 ↓)	42.5 (10.2 ↓)	34.3 (11.3 ↓)	39.0 (9.9 ↓)
w/ PT+FT	75.7 (2.1 ↓)	60.4 (1.9 ↓)	44.2 (5.5 ↓)	58.0 (0.2 ↓)	51.6 (6.8 ↓)	86.9 (1.5 ↓)	37.3 (15.4 ↓)	29.7 (15.9 ↓)	34.3 (14.6 ↓)

Table 5: Ablation studies of Cobra-3.5B on vision encoders, projectors, language models and training strategies



Visual input example	Spatial Reasoning	Scene Description
		
User	Is the bicycle parked to the right of the dog?	What’s going on in this image?
Cobra-3.5B (ours)	The bicycle is actually parked on the left side of the dog, not the right .	A monkey is holding two knives in its hands, while a man in the background is wearing a white shirt with a floral pattern .
LLaVA v1.5-7B	Yes , the bicycle is parked on the right side of the dog.	In this image, a monkey is holding two knives in its hands, seemingly posing for a picture.

Table 6: **Visualization of the MLLM example.** Cobra generates accurate and more detailed textual descriptions compared with the baseline, where **green** indicates a correct answer, **red** produces inaccurate answers and **blue** is a more detailed description.

Projectors. Besides, a different choice of projection layer is used in the experiments. We investigate a lightweight down-sample projector (LDPv2) to see if we can further speed up the inference process without obvious deterioration in performance. Applying LDPv2 to Cobra harms the performance on all benchmarks except VizWiz. Unfortunately, we observed that the models using LDPv2 show a significant decrease in accuracy on TextVQA, VSR, and localization benchmarks, which require precise visual information.

Base or instruct-tuned LLMs. We also explored the application of different Mamba LLMs. Specifically, we chose a base model that had not been fine-tuned on any chat datasets. As indicated in Table 3, the fine-tuned model achieved notably higher accuracy on the VizWiz and VSR benchmarks compared to the pre-trained model that did not utilize chat corpora, with accuracy improvements of 2.5% and 4%, respectively. On other benchmarks, the differences were not significant with accuracy gaps within 1%. The chat model exhibits a slight disadvantage compared to the base model only on the GQA and POPE benchmarks.

Training strategies. Different training strategies were investigated. The results show that fine-tuning the language model for two epochs yields strictly better performance on all evaluated benchmarks compared with the model that only fine-tuned for one epoch. This suggests that the model may be underfitted with only one epoch of training.

Additionally, we discovered that initializing a pre-aligned projector during the fine-tuning stage actually harms the model’s performance, resulting in consistently lower accuracy across all benchmarks compared to a model with a randomly initialized projector (when both models are fine-tuned for one epoch) except TextVQA. This conclusion differs

from several different approaches that treat pre-alignment as the first stage of training. (Lin et al. 2024; Chu et al. 2024).

Visualization of the MLLM example. We visualize some examples to demonstrate the performance. In Table 6, Cobra outperforms LLaVA v1.5 in the first example involving the judgment of spatial relationships. Cobra correctly identified that the dog was parked to the right of the bicycle, whereas LLaVA provided the opposite, incorrect answer. In the second example, Cobra offered a more detailed description of the background information compared with LLaVA. More examples are shown in the supplementary material.

5 Conclusion

In this study, we propose Cobra, which addresses the efficiency bottleneck of existing multi-modal large language models (MLLMs) that rely on Transformer networks with quadratic computational complexity. We explore combining language models with linear computational complexity and multi-modal inputs. In terms of fusing visual and linguistic information, we have successfully optimized the internal information integration of the Mamba language model through in-depth research on different modality fusion schemes, achieving more effective multi-modal representation. Experiments demonstrate that Cobra not only significantly improves computational efficiency, but also performs comparably to advanced models like LLaVA, especially excelling in overcoming visual hallucination and spatial relationship judgment. It even significantly reduces the number of parameters. This opens up new possibilities for deploying high-performance AI models in environments that require high-frequency processing of visual information, such as visual-based robotic feedback control, in the future.

Acknowledgments

This work was supported by the National Science and Technology Innovation 2030 - Major Project (Grant No. 2022ZD0208800), and NSFC General Program (Grant No. 62176215).

References

- Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; et al. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. arXiv:2204.14198.
- Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.
- Awadalla, A.; Gao, I.; Gardner, J.; Hessel, J.; Hanafy, Y.; et al. 2023. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. arXiv:2308.01390.
- Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; et al. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966.
- Bar-David, S.; Zimerman, I.; Nachmani, E.; and Wolf, L. 2023. Decision S4: Efficient Sequence-Based RL via State Spaces Layers. arXiv:2306.05167.
- Bellagente, M.; Tow, J.; Mahan, D.; Phung, D.; Zhuravinskyi, M.; Adithyan, R.; Baicoianu, J.; Brooks, B.; Cooper, N.; Datta, A.; Lee, M.; Mostaque, E.; Pieler, M.; Pinnaparju, N.; Rocha, P.; Saini, H.; Teufel, H.; Zanichelli, N.; and Riquelme, C. 2024. Stable LM 2 1.6B Technical Report. arXiv:2402.17834.
- Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Chen, X.; et al. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. arXiv:2307.15818.
- Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; et al. 2023a. Shikra: Unleashing Multimodal LLM’s Referential Dialogue Magic. arXiv:2306.15195.
- Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; et al. 2023b. ShareGPT4V: Improving Large Multi-Modal Models with Better Captions. arXiv:2311.12793.
- Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.
- Chu, X.; Qiao, L.; Lin, X.; Xu, S.; Yang, Y.; Hu, Y.; et al. 2023. MobileVLM : A Fast, Strong and Open Vision Language Assistant for Mobile Devices. arXiv:2312.16886.
- Chu, X.; Qiao, L.; Zhang, X.; Xu, S.; Wei, F.; Yang, Y.; et al. 2024. MobileVLM V2: Faster and Stronger Baseline for Vision Language Model. arXiv:2402.03766.
- Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; et al. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500.
- Dao, T.; and Gu, A. 2024. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. arXiv:2405.21060.
- Ding, P.; Zhao, H.; Song, W.; Zhang, W.; Zhang, M.; et al. 2024. QUAR-VLA: Vision-Language-Action Model for Quadruped Robots. arXiv:2312.14457.
- Gao, P.; Han, J.; Zhang, R.; Lin, Z.; Geng, S.; et al. 2023. LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model. arXiv:2304.15010.
- 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. arXiv:1612.00837.
- Gu, A.; and Dao, T. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv:2312.00752.
- Gu, A.; Goel, K.; and Ré, C. 2022. Efficiently Modeling Long Sequences with Structured State Spaces. arXiv:2111.00396.
- Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; et al. 2018. VizWiz Grand Challenge: Answering Visual Questions from Blind People. arXiv:1802.08218.
- Hasani, R.; Lechner, M.; Wang, T.-H.; Chahine, M.; Amini, A.; and Rus, D. 2022. Liquid Structural State-Space Models. arXiv:2209.12951.
- Hendriksen, M.; Vakulenko, S.; Kuiper, E.; and de Rijke, M. 2023. Scene-centric vs. object-centric image-text cross-modal retrieval: a reproducibility study. In *European Conference on Information Retrieval*, 68–85. Springer.
- Hudson, D. A.; and Manning, C. D. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. arXiv:1902.09506.
- Karamcheti, S.; Nair, S.; Balakrishna, A.; Liang, P.; Kollar, T.; et al. 2024. Prismatic VLMs: Investigating the Design Space of Visually-Conditioned Language Models. arXiv:2402.07865.
- Katharopoulos, A.; Vyas, A.; Pappas, N.; and Fleuret, F. 2020. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. arXiv:2006.16236.
- Kazemzadeh, S.; Ordonez, V.; andre Matten, M.; and Berg, T. L. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Conference on Empirical Methods in Natural Language Processing*.
- Ke, L.; Pei, W.; Li, R.; Shen, X.; and Tai, Y.-W. 2019. Reflective decoding network for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8888–8897.
- Kim, M. J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; et al. 2024. OpenVLA: An Open-Source Vision-Language-Action Model. arXiv:2406.09246.
- Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; et al. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. arXiv:1602.07332.
- Laurençon, H.; Saulnier, L.; Tronchon, L.; Bekman, S.; Singh, A.; et al. 2023. OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. arXiv:2306.16527.

- 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. arXiv:2301.12597.
- Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; et al. 2023b. Evaluating Object Hallucination in Large Vision-Language Models. arXiv:2305.10355.
- Lin, B.; Tang, Z.; Ye, Y.; Cui, J.; Zhu, B.; Jin, P.; et al. 2024. MoE-LLaVA: Mixture of Experts for Large Vision-Language Models. arXiv:2401.15947.
- Liu, F.; Emerson, G.; and Collier, N. 2023. Visual Spatial Reasoning. arXiv:2205.00363.
- Liu, F.; Lin, K.; Li, L.; Wang, J.; Yacoob, Y.; et al. 2023a. Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. arXiv:2306.14565.
- Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023b. Improved Baselines with Visual Instruction Tuning. arXiv:2310.03744.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023c. Visual Instruction Tuning. arXiv:2304.08485.
- Lu, C.; Schroecker, Y.; Gu, A.; Parisotto, E.; Foerster, J.; et al. 2023. Structured State Space Models for In-Context Reinforcement Learning. arXiv:2303.03982.
- Lu, J.; Clark, C.; Zellers, R.; Mottaghi, R.; and Kembhavi, A. 2022. Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks. arXiv:2206.08916.
- Mercat, J.; Vasiljevic, I.; Keh, S.; Arora, K.; Dave, A.; Gaidon, A.; and Kollar, T. 2024. Linearizing Large Language Models. arXiv:2405.06640.
- OpenAI. 2023. GPT-4V(ision) System Card.
- Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; et al. 2024. DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193.
- ShareGPT. 2023. <https://sharegpt.com/>.
- Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; et al. 2019. Towards VQA Models That Can Read. arXiv:1904.08920.
- Smith, J. T. H.; Warrington, A.; and Linderman, S. W. 2023. Simplified State Space Layers for Sequence Modeling. arXiv:2208.04933.
- Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.
- Tong, S.; Liu, Z.; Zhai, Y.; Ma, Y.; LeCun, Y.; and Xie, S. 2024. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. arXiv:2401.06209.
- Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.
- Waleffe, R.; Byeon, W.; Riach, D.; Norick, B.; Korthikanti, V.; et al. 2024. An Empirical Study of Mamba-based Language Models. arXiv:2406.07887.
- Wang, J.; Meng, L.; Weng, Z.; He, B.; Wu, Z.; et al. 2023. To See is to Believe: Prompting GPT-4V for Better Visual Instruction Tuning. arXiv:2311.07574.
- Wu, S.; Fei, H.; Qu, L.; Ji, W.; and Chua, T.-S. 2024. NExT-GPT: Any-to-Any Multimodal LLM. arXiv:2309.05519.
- Yan, J. N.; Gu, J.; and Rush, A. M. 2023. Diffusion Models Without Attention. arXiv:2311.18257.
- Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling Context in Referring Expressions. arXiv:1608.00272.
- Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid Loss for Language Image Pre-Training. arXiv:2303.15343.
- Zhang, B.; and Sennrich, R. 2019. Root Mean Square Layer Normalization. arXiv:1910.07467.
- Zhang, P.; Zeng, G.; Wang, T.; and Lu, W. 2024. TinyLlama: An Open-Source Small Language Model. arXiv:2401.02385.
- Zhou, B.; Hu, Y.; Weng, X.; Jia, J.; Luo, J.; Liu, X.; et al. 2024. TinyLLaVA: A Framework of Small-scale Large Multimodal Models. arXiv:2402.14289.
- Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv:2304.10592.
- Zhu, Y.; Zhu, M.; Liu, N.; Ou, Z.; Mou, X.; and Tang, J. 2024. LLaVA-Phi: Efficient Multi-Modal Assistant with Small Language Model. arXiv:2401.02330.