

---

# Mini-Monkey: Multi-Scale Adaptive Cropping for Multimodal Large Language Models

---

Mingxin Huang   Yuliang Liu   Dingkang Liang   Lianwen Jin\*   Xiang Bai\*  
 yliu@hust.edu.cn

## Abstract

Recently, there has been significant interest in enhancing the capability of multimodal large language models (MLLMs) to process high-resolution images. Most existing methods focus on adopting a cropping strategy to improve the ability of multimodal large language models to understand image details. However, this cropping operation inevitably causes the segmentation of objects and connected areas, which impairs the MLLM’s ability to recognize small or irregularly shaped objects or text. This issue is particularly evident in lightweight MLLMs. Addressing this issue, we propose Mini-Monkey, a lightweight MLLM that incorporates a plug-and-play method called multi-scale adaptive cropping strategy (MSAC). Mini-Monkey adaptively generates multi-scale representations, allowing it to select non-segmented objects from various scales. To mitigate the computational overhead introduced by MSAC, we propose a Scale Compression Mechanism (SCM), which effectively compresses image tokens. Mini-Monkey achieves state-of-the-art performance among 2B-parameter MLLMs. It not only demonstrates leading performance on a variety of general multimodal understanding tasks but also shows consistent improvements in document understanding capabilities. On the OCRBench, Mini-Monkey achieves a score of 802, outperforming 8B-parameter state-of-the-art model InternVL2-8B. Besides, our model and training strategy are very efficient, which can be trained with only eight RTX 3090. The code is available at <https://github.com/Yuliang-Liu/Monkey>.

## 1 Introduction

In recent years, the field of natural language processing (NLP) has demonstrated a significant paradigm shift, marked by a focus on the development of Large Language Models [80, 3, 66, 56] (LLMs). This shift has paved the way for the creation of multimodal large language models (MLLMs) capable of processing general vision-and-language understanding [33, 41, 2]. Researchers are actively exploring effective and efficient methods for integrating vision encoders with LLMs. Some methods, such as Flamingo [1], BLIP-2 [33], MiniGPT4 [82], and Qwen-VL [2] utilize a set of learnable queries to sample the image tokens and align the image tokens with Large Language Models. In contrast, other methods like LLaVA [42] and CogVLM [67] propose to use a linear layer to achieve this. Despite these achievements, detailed scene understanding was not achieved well by previous multimodal large language models due to the limited resolution to handle.

Recent efforts have attempted to tackle this issue by expanding the input resolution of the image. The cropping strategy is one of the most commonly used methods [40, 74, 36, 8, 60, 72]. There are many technical extensions to the simplest cropping strategy. For instance, Monkey [36] leverages

---

Y. Liu, D. Liang, and X. Bai are with Huazhong University of Science and Technology. M. Huang and L. Jin are with South China University of Technology. This work was done when M. Huang was visiting Huazhong University of Science and Technology. \*Corresponding authors.

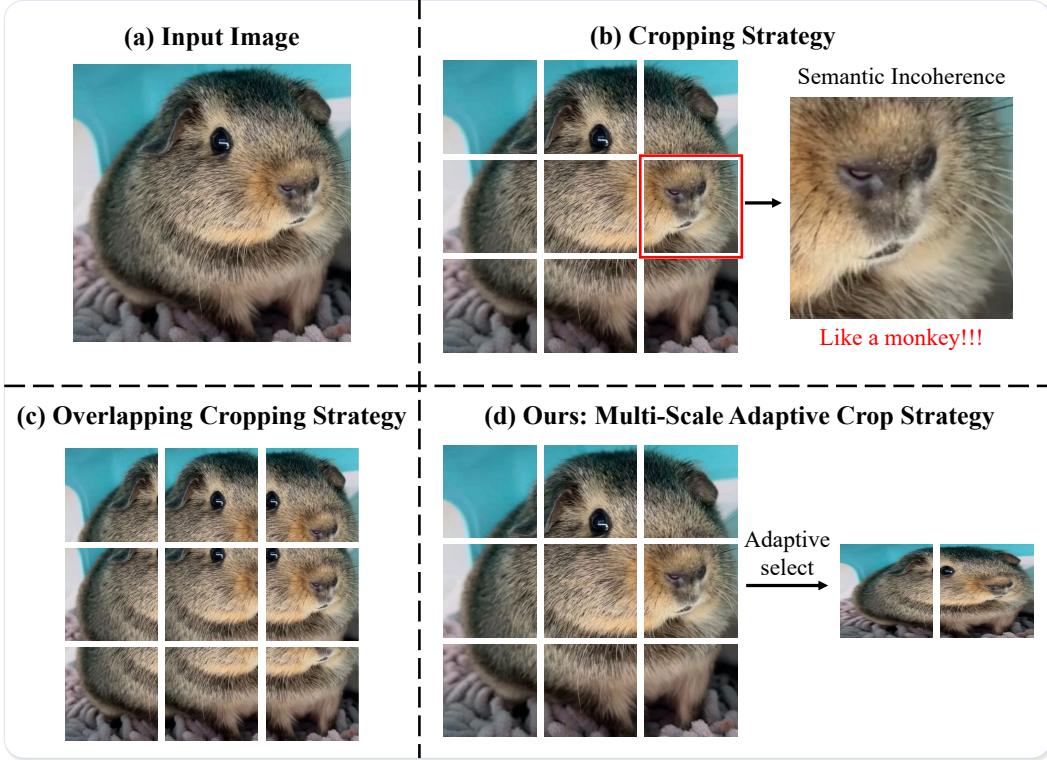


Figure 1: Sawtooth Effect caused by the cropping. (a) Input Image (b) Cropping strategy. (c) Overlapping Cropping Strategy. (d) Ours: Multi-scale adaptive cropping strategy.

the LoRA [23] into the vision encoder for learning detail-sensitive features from the sub-image. Although these methods have shown promising results, their performance still lags behind leading commercial models. To bridge this gap, InternVL 1.5 [8] employs a powerful vision encoder [9] to enhance visual representation and uses dynamic high resolution to scale up the resolution to 4K, significantly improving the performance.

Despite the significant progress achieved by multimodal large language models, challenges in detailed scene understanding persist due to the cropping strategy. Cropping operations on images inevitably segment objects and connected areas, impairing the MLLM’s ability to recognize small or irregularly shaped objects, particularly in the context of document understanding. This strategy will introduce two types of semantic incoherence: 1) If an object or character is divided, it may not be recognized [25]. For instance, after cropping, the nose looks very much like a monkey, as shown in Fig. 1(b); 2) If a word or sentence is segmented, the semantic damage of the segmented word will be caused. For example, the word ‘Breakdown’ may be divided into ‘Break’ and ‘down’, causing semantic damage to the segmented word [45, 79]. For simplicity, we call this issue the sawtooth effect in this paper. A very straightforward idea is to adopt an overlapping cropping strategy to solve this issue, as presented in Fig. 1(c). However, as presented in our ablation studies Sec. 4.3, the overlapping cropping strategy introduces certain hallucinations that cause performance to decrease rather than increase. Moreover, this sawtooth effect is particularly evident in lightweight MLLMs, as discussed in Sec. 4.4. Larger MLLMs with enhanced comprehension capabilities can alleviate this issue to some extent.

In this paper, we propose Mini-Monkey, a lightweight multimodal large language model designed to mitigate the sawtooth effect caused by cropping strategies. Unlike existing methods [74, 8, 37] that directly crop input images, Mini-Monkey employs a plug-and-play method termed multi-scale adaptive cropping strategy (MSAC), which enables effective complementation between features from different scales, as shown in Fig. 1(d). MSAC first performs a stratified operation on a pre-set group of grids according to the aspect ratios and the resolution of these grids. It then adaptively selects multiple aspect ratios from each stratified layer, ensuring that the same text is not split across different images. Multiple images will be generated based on the aspect ratios and processed by a pre-trained

vision encoder to generate multi-scale visual representations. These representations are concatenated into a sequence and fused within the LLM to interact with each other. With the MSAC, Mini-Monkey adaptively generates multi-scale representations, allowing the model to select non-segmented object features from various scales. The MSAC may introduce some additional computational overhead. Therefore, we propose a Scale Compression Mechanism (SCM) for use in situations where there are restrictions on computational overhead. SCM is a training-free and parameter-free module to reduce the computational overhead. It utilizes the well-trained attention layers from the LLM to produce the attention weight and dropout token based on the attention weight.

Experiments have demonstrated the effectiveness of Mini-Monkey: 1) Mini-Monkey achieves state-of-the-art performance among 2B-parameter MLLMs in both general multimodal understanding and document understanding tasks. Especially, Mini-Monkey outperforms the state-of-the-art 2B-parameter method by an average of 1.7% across 13 benchmarks in terms of evaluation metrics; 2) Surprisingly, we find that Mini-Monkey achieves a score of 802 on the OCRBench, outperforming the 8B-parameter state-of-the-art model InternVL2-8B. Additionally, the training of Mini-Monkey is efficient that our method can be trained using only eight RTX 3090.

## 2 Related Works

### 2.1 Multimodal Large Language Models

In recent years, Large Language Models (LLMs) have made significant progress [80, 3, 66, 56? ]. Drawing from this advancement, many efforts have been made to integrate a vision encoder into Large Language Models for vision-language understanding. A commonly employed approach is the Linear Projector method[41, 67, 82], which maps the output of the vision encoder to the same feature space as the text features of the Large Language Models. Some methods, such as QFormer [33], Perceiver Resampler [1, 2], or Abstractor [75], introduce a set of learnable queries to facilitate this integration. Despite notable progress, previous methods face challenges in handling the detailed scene understanding due to the limited resolution. To address this issue, recent works have primarily employed the following strategies: 1) Two vision encoders, one for processing high-resolution images and one for processing low-resolution images [69, 81, 21]. 2) Directly using visual encoders that support high-resolution input [38, 50]. 3) Using a cropping strategy to segment the high-resolution images into several low-resolution images [74, 37]. While these methods have effectively enhanced resolution, they still display shortcomings in document understanding, especially when compared to top commercial models. To close the gap, InternVL1.5 [8] utilizes a large vision encoder [9] and a dynamic high-resolution strategy to train on high-quality data. Concurrently, LLama3-V [65] employed a cropping strategy to enhance resolution, releasing several models with varying parameter counts, reaching up to 400 billion. Although LLama3-V and InternVL1.5 achieve promising results on several multimodal benchmarks, the cropping strategy used in it will inevitably result in semantic incoherence: 1) If an object or character is divided, it may not be recognized; 2) If the word or sentence is segmented, the semantic damage of the segmented word will be caused. For example, the word ‘Breakdown’ may be divided into ‘Break’ and ‘down’, causing semantic damage to the segmented word [45, 79]. This will limit it in the detailed scene understanding. Although some methods [45, 25] attempt to address this issue by introducing attention modules, they introduce additional parameters and require training this module from scratch. In contrast, our method is plug-and-play, requiring no additional parameters.

### 2.2 Visually-Situated Document Understanding

The visually-situated document understanding is a task that comprehends rich text information in the images, including natural images [61, 62], documents images [55, 63, 53], charts images [52, 27], tables images [57, 7], etc. document understanding models can be broadly categorized into two types based on their reliance on OCR systems for text extraction: OCR-dependent methods and OCR-free methods. OCR-dependent methods use the text extracted from the OCR system to perform related document understanding tasks. For instance, LayoutLM v3 [26] learns the multimodal representations by unified text and image masking pre-training objectives. UDOP [64] develops a unified framework to learn and generate vision, text, and layout modalities together. On the contrary, OCR-free methods perform the document understanding tasks in an end-to-end manner without OCR input. Dount [29] directly generates textual elements based on the document images without OCR input. Pix2Struct [31]

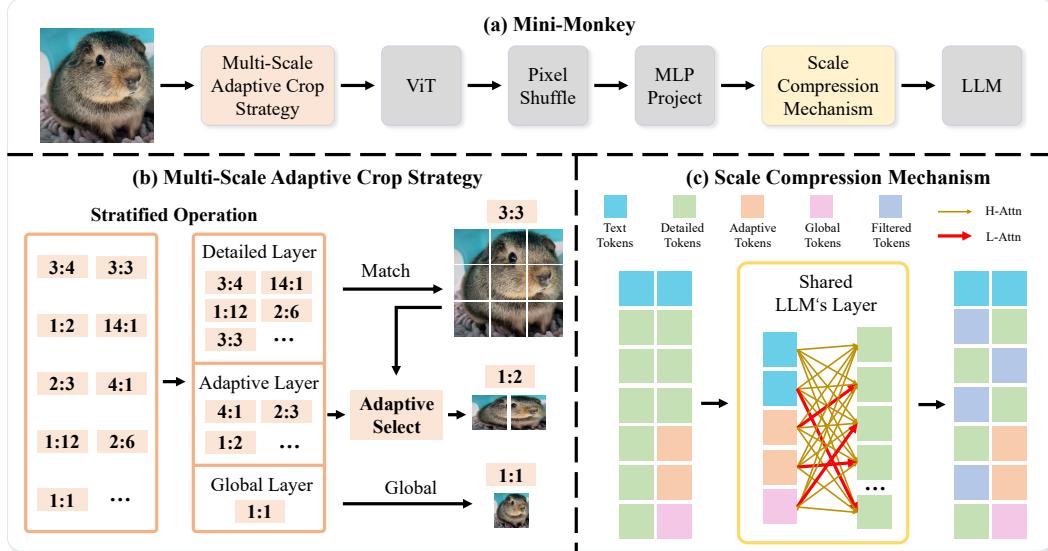


Figure 2: The overall architecture of Mini-Monkey. H-Attn represents high attention weight. L-Attn represents low attention weights. The tokens with low attention weights will be filtered. The shared LLM’s Layer represents using the block layer from LLM in SCM.

further leverages structure information by generating the HTML DOM tree for masked screenshots of web pages. Additionally, some methods incorporate large language models (LLMs) for enhanced document understanding. Ureader [74] presents a shape-adaptive cropping module to crop input images based on their aspect ratios. Similarly, TextMonkey [45] introduces enhancements for cross-window relations and a token resampler for document understanding. LayoutLLM [50] develops LayoutCoT for layout-aware supervised fine-tuning. StrucTexTv3 [51] introduces a lightweight multimodal large language model designed for perceiving and comprehending text-rich images.

### 2.3 Lightweight Multimodal Large Language Models

Due to the substantial computational costs associated with multimodal large language models (MLLMs), recent efforts have focused on developing more efficient models for rapid development and real-world applications. LLaVA-Phi [83] and Imp [59] leverage a lightweight language model combined with a vision encoder to create a lightweight large multimodal model. In this context, several researchers are exploring efficient architectural designs. For instance, MobileVLM [10] introduces a lightweight downsample projector to minimize resource usage during training and inference. Bunny [20] offers an efficient data compression technique to reduce the volume of pretraining data required. TinyGPT-V [78] employs a multi-stage training approach tailored for lightweight multimodal models. Similarly, MiniCPM presents a scalable training strategy aimed at producing an efficient lightweight multimodal large language model. Additionally, Vary-toy [70] supports high-resolution input, while InternVL 2 [8] enhances the performance of lightweight MLLMs through a dynamic high-resolution strategy. Despite these promising advancements, the state-of-the-art method faces limitations due to a sawtooth effect caused by the cropping strategy.

## 3 Mini-Monkey

The overall architecture is illustrated in Fig. 2. Mini-Monkey consists of an multi-scale adaptive cropping strategy (MSAC), a vision encoder, an MLP layer, a Scale Compression Mechanism (SCM), and a Large Language Model (LLM). Initially, Mini-Monkey generates multiple images through MSAC. These images are then processed by the vision encoder and MLP layer to extract image tokens. The Scale Compression Mechanism adjusts these image tokens based on the input question and forwards them to the LLM, which subsequently generates the final answers.

### 3.1 Multi-Scale Adaptive Cropping Strategy

Previous state-of-the-art methods [37, 8] adopt the cropping-based strategy to expand the resolution of images and segment the high-resolution images into a set of sub-images. However, this cropping strategy will lead to a sawtooth effect. To address this issue, we introduce a multi-scale adaptive cropping strategy (MSAC) that achieves the synergy between images with different scales to mitigate the semantic incoherence caused by the cropping strategy. As shown in Fig. 2 (b), we generate a pre-defined set of grids. The maximum of these grids is less than  $Max_{num}$ .

Then, we perform a stratified operation on these grids, which are divided into three sets according to their aspect ratios. We will select one aspect ratio for each layer. Different stratified layers provide different information to the model. The detailed layer  $A_d$  is responsible for providing detailed information. It not only limits the maximum of the sub-image but also limits the minimum of the sub-image to make the image as large as possible to make the object in the image clearer. Due to the cropping strategy, the images generated by this layer may have a semantic inconsistency. Therefore, we utilize the adaptive layer  $A_a$  to synergize with the detailed layer, allowing the model to select non-segmented objects from various scales. The adaptive layer will adaptively generate an aspect ratio based on the detailed layer, ensuring that the cropping lines on the detailed layer and those on the adaptive layer do not overlap. This can be formulated as follows:

$$C_{Id} \cap C_{Ia} = \emptyset. \quad (1)$$

where  $C_{Id}$  represents the cropping lines on the detailed layer.  $C_{Ia}$  represents the cropping lines on the adaptive layer. Specifically, if the aspect ratio from the detailed layer is a multiple of that from the adaptive layer, we remove it from the adaptive layer and select a new ratio. This process ensures that the detailed and adaptive layers provide distinct semantic information and visual features for the model.

We also produce a global view of the image as a low resolution using an aspect ratio 1 : 1, termed global layer. After obtaining the image from three layers, these images are sent to the vision encoder to extract the features and compress the visual tokens through the scale compression mechanism. Once the visual tokens are compressed, they will be fed into the large language model to conduct the multi-scale visual representations fusion and output the results.

**Multi-Scale Visual Representations Fusion.** Unlike the previous work [60] that simply concatenates multi-scale features along the dimension, our approach involves the fusion of multi-scale visual representations within a large language model (LLM). Within the LLM, these multi-scale visual representations interact with each other through self-attention. By fusing features from different scales, Mini-Monkey gains an enhanced ability to comprehend visual text information.

### 3.2 Scale Compression Mechanism

Although the proposed MSAC significantly enhances model performance, certain scenarios may impose computational requirements. To tackle this challenge, we introduce a parameter-free token compression method called the Scale Compression Mechanism (SCM), which is used to reduce the visual tokens, as shown in Fig. 2 (c). Due to the lower information density of tokens from detailed layers, we primarily focus on compressing these tokens. In contrast, visual tokens from adaptive and global layers provide the LLM with complete spatial information. Specifically, a well-trained LLM from MLLM can effectively select the necessary visual features based on the input question. Consequently, SCM utilizes the first and second layers of the LLM to select visual tokens without generating any additional parameters. The input visual token including  $V_d \in \mathbb{R}^{L_1 \times C}$ ,  $V_a \in \mathbb{R}^{L_2 \times C}$ , and  $V_g \in \mathbb{R}^{L_3 \times C}$ , and the textual token  $T_t \in \mathbb{R}^{T \times C}$  will be sent into an LLM's Layer.  $V_d$  represents the tokens from the detailed layer.  $V_a$  represents the tokens from adaptive layer.  $V_g$  represents the tokens from the global layer. Notable, we reuse the layer of the LLM as this LLM's Layer. The LLM's Layer will output an attention map. We choose the visual token from the adaptive layer, global, and textual token to attend to the visual token from the detailed layer. The calculation of the attention can be formulated as follows:

$$\text{Attn}_w = \text{softmax}\left(\frac{(Cat(V_a, V_g, T_t) + PE(Cat(V_a, V_g, T_t)))(V_d + PE(V_d))^T}{\sqrt{D}}\right). \quad (2)$$

where PE represents the position encoding and  $D$  denotes the dimension of the LLM.  $Cat()$  represents the sequence concatenation operation. After computing the attention mechanism, we average the

Table 1: Comparison with SoTA models on 16 multimodal benchmarks. General multimodal benchmarks encompass: MME [17], RealWorldQA [71], AI2D test [28], CCBench [43], SEED Image [32], HallusionBench [19], and POPE [35]. Additionally, the math dataset includes MathVista testmini [49]. The MME results we report are the sum of the perception and cognition scores.  $\dagger$  represents the results from the OpenCompass leaderboard [11].

model	#param	General Multimodal Benchmarks						Math MathVista
		MME	RWQA	AI2D	CCB	SEED	HallB	
Mini-Gemini [34]	35B	2141.0	—	—	—	—	—	43.3
LLaVA-NeXT [40]	35B	2028.0	—	74.9	49.2	75.9	34.8	<b>89.6</b> <sup>§</sup>
InternVL 1.2 [9]	40B	2175.4	<b>67.5</b>	79.0	59.2	75.6	47.6	88.0
InternVL 1.5 [8]	26B	<b>2187.8</b>	66.0	<b>80.7</b>	<b>69.8</b>	<b>76.0</b>	<b>49.3</b>	88.3
DeepSeek-VL [48]	1.7B	1531.6	49.7 <sup>§</sup>	51.5 <sup>§</sup>	37.6 <sup>§</sup>	43.7 <sup>§</sup>	27.6 <sup>§</sup>	85.9 <sup>§</sup>
Mini-Gemini [34]	2.2B	1653.0	-	-	-	-	-	29.4
Bunny-StableLM-2 [20]	2B	1602.9	-	-	-	58.8	-	85.9
MiniCPM-V-2 [73]	2.8B	1808.6	55.8 <sup>§</sup>	62.9 <sup>§</sup>	48.0 <sup>§</sup>	-	36.1 <sup>§</sup>	86.3 <sup>§</sup>
InternVL 2 [8]	2B	1876.8	57.3	74.1	74.7	70.9 <sup>§</sup>	37.9	85.2 <sup>§</sup>
Mini-Monkey (ours)	2B	<b>1881.9</b>	<b>57.5</b>	<b>74.7</b>	<b>75.5</b>	<b>71.3</b>	<b>38.7</b>	<b>86.7</b>
								<b>47.3</b>

first dimension of the attention map  $\text{Attn}_w \in \mathbb{R}^{(L_2+L_3+T) \times L_1}$  to obtain a weight vector  $\mathbf{W}_a \in \mathbb{R}^{L_1}$ . Subsequently, we select the top  $K$  visual features from detailed layers based on this weight vector  $\mathbf{W}_a$ . These selected tokens, along with tokens from the adaptive layer, global layer, and the textual token, are input into the LLM to generate the results. Compared to FastV [6], SCM is more targeted by using tokens with high relative information density to compress tokens with low information density. The ablation study in Sec. 4.3 demonstrates the effectiveness of SCM.

## 4 Experiments

### 4.1 Implementation Details

We use a well-trained InternViT [9], MLP layers, and the InternLLM [4] from InternVL2-2B [8] as the vision encoder, connector and the LLM. Following previous work [9], we use the (448, 448) as the input resolution of InternViT. The training datasets used to train the model include DocVQA [55], ChartQA [52], DVQA [27], AI2D [28], GeoQA+ [5], and LLaVA-150K (zh) [41]. We use the AdamW [47] as the optimizer. The base learning rate is 4e-8.

**Evaluation.** Following the previous work [20, 8], we evaluate Mini-Monkey on eleven general multimodal understanding benchmarks, including MathVista testmini [49], SEED Image [32], RealWorldQA [71], AI2D test [28], POPE [35], CCBench [43], MME [17], and HallusionBench [19].

For document understanding, following the previous work [45], we employ two distinct types of metrics to assess the performance of Mini-Monkey. Initially, we leverage the standard metrics provided by the benchmarks to evaluate Mini-Monkey. For this metric, similar to [8], we utilize benchmarks such as ChartQA [52], DocVQA [55], InfoVQA [54], TextVQA [62], and OCRBench [44]. ChartQA, DocVQA, InfoVQA, and TextVQA are widely used to assess the textual comprehension of models. OCRBench, a more recent benchmark, includes 29 datasets to provide a comprehensive evaluation of the model’s capabilities. Subsequently, we apply the accuracy metric to verify the performance. For this metric, a response from Mini-Monkey that fully captures the ground truth is considered a true positive. Further details on this metric and the used benchmarks can be referenced in [44].

### 4.2 Comparison to the State of the Art

**General Multimodal Understanding.** We evaluate Mini-Monkey on general multimodal understanding following [20, 8]. The results are shown in 1. Mini-Monkey surpasses other 2B-parameter models on 11 benchmarks. On the MathVista and POIE, Mini-Monkey outperforms the previous state-of-the-art method InternVL2-2B by 1% and 1.5%, respectively. On the HallusionBench, Mini-Monkey outperforms MiniCPM-V-2 by 2.6%. These results showcase the ability of Mini-Monkey to handle general multimodal understanding and reasoning tasks.

Table 2: Comparison to state-of-the-art MLLMs on OCR-related Tasks. Mini-Monkey achieves the best results among the 2B-parameter MLLMs.  $\dagger$  represents the results from the OpenCompass leaderboard [11].

Model	Model Size	$\text{DocVQA}^{\text{Test}}$	$\text{ChartQA}^{\text{Test}}$	$\text{InfoVQA}^{\text{Test}}$	$\text{TextVQA}^{\text{Val}}$	OCRBench
TextMonkey [45]	9B	73.0	66.9	28.6	65.6	558
TextHawk [77]	7B	76.4	66.6	50.6	—	—
DocKylin [79]	7B	77.3	46.6	66.8	—	—
HiRes-LLaVA [25]	7B	74.7	61.5	48.0	65.4	—
LLava-UHD [72]	13B	—	—	—	67.7	—
CogAgent [21]	17B	81.6	68.4	44.5	76.1	590
UReader [74]	7B	65.4	59.3	42.2	57.6	—
DocOwl 1.5 [22]	8B	82.2	70.2	50.7	68.6	—
HRVDA [38]	7B	72.1	67.6	43.5	—	—
IXC2-4KHD [14]	8B	90.0	81.0	68.6	77.2	675
InternVL 1.5 [8]	26B	90.9	<b>83.8</b>	72.5	<b>80.6</b>	724
InternVL 2 [8]	8B	<b>91.6</b>	83.3	<b>74.8</b>	77.4	<b>794</b>
GLM4-V [18]	9B	-	-	-	-	786
Vary-toy [70]	1.8B	65.6	59.1	-	-	-
MiniCPM-V 2.0 [73]	2.8B	71.9	55.6 $\dagger$	-	74.1	605
InternVL 2 [8]	2B	86.9	76.2	58.9	73.4	784
Mini-Monkey (Ours)	2B	<b>87.4</b>	<b>76.5</b>	<b>60.1</b>	<b>75.7</b>	<b>802</b>

Table 3: Quantitative accuracy (%) comparison of our model with existing multimodal large language models (MLLMs) on several benchmarks. Following TextMonkey [45], we use the accuracy metrics to evaluate our method.

Method	Scene Text-Centric VQA		Document-Oriented VQA			FUNSD	KIE	
	STVQA	TextVQA	DocVQA	InfoVQA	ChartQA		SROIE	POIE
BLIP2-OPT-6.7B [33]	20.9	23.5	3.2	11.3	3.4	0.2	0.1	0.3
mPLUG-Owl [75]	30.5	34.0	7.4	20.0	7.9	0.5	1.7	2.5
InstructBLIP [12]	27.4	29.1	4.5	16.4	5.3	0.2	0.6	1.0
LLaVAR [81]	39.2	41.8	12.3	16.5	12.2	0.5	5.2	5.9
BLIVA [24]	32.1	33.3	5.8	23.6	8.7	0.2	0.7	2.1
mPLUG-Owl2-8 [76]	49.8	53.9	17.9	18.9	19.4	1.4	3.2	9.9
LLaVA1.5-7B [39]	38.1	38.7	8.5	14.7	9.3	0.2	1.7	2.5
TGDoc [68]	36.3	46.2	9.0	12.8	12.7	1.4	3.0	22.2
UniDoc [16]	35.2	46.2	7.7	14.7	10.9	1.0	2.9	5.1
DocPedia [15]	45.5	60.2	47.1	15.2	46.9	29.9	21.4	39.9
Monkey-8B [37]	54.7	64.3	50.1	25.8	54.0	24.1	41.9	19.9
InternVL-8B [9]	62.2	59.8	28.7	23.6	45.6	6.5	26.4	25.9
InternLM-XComposer2-7B [13]	59.6	62.2	39.7	28.6	51.6	15.3	34.2	49.3
TextMonkey-9B [45]	61.8	65.9	64.3	28.2	<b>58.2</b>	32.3	47.0	27.9
Mini-Monkey-2B (Ours)	<b>66.5</b>	<b>68.4</b>	<b>78.1</b>	<b>49.6</b>	57.9	<b>42.9</b>	<b>70.3</b>	<b>69.9</b>

Table 4: Ablation study of multi-scale adaptive cropping strategy. We compare our method with the existing cropping strategy and the overlay cropping strategy.

Model	Resolution Strategy	TextVQA	OCRBench	MME	HallIB	POPE
Baseline	Dynamic High-Resolution strategy [8]	73.4	784	1876.8	37.9	85.2
Baseline	Fixed Size High-Resolution strategy [37]	74.2	772	1824.5	37.6	85.0
Baseline	Overlapping Cropping Strategy	70.6	758	1874.1	36.8	83.5
Baseline	Multi-Scale Strategy [60]	74.8	776	1846.8	38.1	85.3
Mini-Monkey (Ours)	Multi-Scale Adaptive cropping strategy	<b>75.7</b>	<b>802</b>	<b>1881.9</b>	<b>38.7</b>	<b>86.7</b>

**Document Understanding.** For the first type of metric, the results are presented in Tab. 2. We use a relaxed accuracy measure for ChartQA, ANLS for DocVQA and InfoVQA, and the VQA score for TextVQA. The results indicate that Mini-Monkey achieves state-of-the-art performance among 2B-parameter multimodal large language models. Compared to InternVL2-2B, our method outperforms it by 2.3%, 1.8%, and 1.2% for TextVQA, InfoVQA, and OCRBench, respectively. Due to the small original resolution of ChartVQA, it is less impacted by cropping operations, resulting in a minor improvement from our method. Notably, in the OCRBench, Mini-Monkey even surpasses the 8B-parameter Large Multimodal Model InternVL2-8B and the 9B-parameter Large Multimodal

Table 5: Ablation study of incorporating multi-scale adaptive cropping strategy to other MLLMs. MSAC represents the multi-scale adaptive cropping strategy.  $\dagger$  represents the results from the OpenCompass leaderboard [11].

Model	MSAC	TextVQA	OCRBench	MME	HallB	POPE
MiniCPM-V-2	$\times$	74.1	605	1808.6	36.1 $\dagger$	86.3 $\dagger$
MiniCPM-V-2	$\checkmark$	76.0 ( <b>+1.9</b> )	627 ( <b>+22</b> )	1819.5 ( <b>+10.9</b> )	36.5 ( <b>+0.4</b> )	87.1 ( <b>+0.8</b> )
InternVL 2	$\times$	73.4	784	1876.8	37.9	85.2
InternVL 2	$\checkmark$	75.7 ( <b>+2.3</b> )	802 ( <b>+18</b> )	1881.9 ( <b>+5.1</b> )	38.7 ( <b>+0.8</b> )	86.7 ( <b>+1.5</b> )

Table 6: Ablation study of the scale compression mechanism. We used different compression ratios to compare with FastV [6]. (0.5) represents 50% compression and (0.9) represents 90% compression.

Model	Resolution Strategy	TextVQA	OCRBench	MME	HallB	POPE
Mini-Monkey	Pooling (0.5)	47.6	256	1765.2	31.5	84.5
Mini-Monkey	Random (0.5)	63.5	503	1805.5	36.2	85.9
Mini-Monkey	FastV [6] (0.5)	73.4	781	1848.0	38.3	83.9
Mini-Monkey	FastV [6] (0.9)	73.9	792	1866.1	37.5	85.8
Mini-Monkey	SCM (0.5)	74.7	794	<b>1886.0</b>	<b>38.7</b>	86.1
Mini-Monkey	SCM (0.9)	<b>75.2</b>	<b>801</b>	1884.7	38.6	<b>86.2</b>

Model GLM4-V by 0.8% and 1.6%, respectively. These results demonstrate the advantages of a multi-scale adaptive cropping strategy in enhancing document understanding.

For the accuracy metric, the results are shown in Tab. 3. Mini-Monkey shows an average performance improvement of 14.8% compared to TextMonkey-9B [45], demonstrating the effectiveness of our method. Mini-Monkey also outperforms the state-of-the-art methods on multiple text-related benchmarks. Specifically, Mini-Monkey achieves 49.2% on FUNSD, 70.3% on SROIE, and 69.9% on POIE, outperforming the previous state-of-the-art method by 10.6%, 23.3%, and 42.0%, respectively. These results further indicate the great potential of Mini-Monkey for downstream task applications, such as visual key information extraction.

### 4.3 Ablation Study

In this section, we perform ablation studies on both general multimodal understanding and document understanding benchmarks to validate the effectiveness of our method. We adopt the TextVQA [62], OCRBench [44], HallusionBench [19], MME [17], and RealWorldQA [71] to conduct ablation studies.

**Multi-Scale Adaptive Cropping Strategy.** We conducted ablation studies to investigate the effectiveness of the proposed multi-scale adaptive cropping strategy. We compared our method with several alternatives: The dynamic high-resolution strategy [8], which maintains aspect ratios to increase resolution. The fixed-size high-resolution strategy [37], which uses a fixed size to increase resolution. The overlapping cropping strategy, which uses a high-resolution approach but crops with overlay. The multi-scale strategy [60], which introduce a fixed-size multi-scale strategy to the MLLM. As presented in Tab. 4, the proposed multi-scale adaptive cropping strategy achieved the best results. Our method outperforms the fixed-size multi-scale strategy [60] by 3.3%. Notably, the over-overlay cropping strategy, instead of improving the model’s performance, actually degraded it.

The proposed multi-scale adaptive crop (MSAC) strategy can be seamlessly integrated into crop-based methods. To demonstrate its effectiveness, we incorporated MSAC into various structures of MLLM, such as MiniCPM-V-2 and InternVL 2. As shown in Tab. 5, MSAC consistently enhances the performance across different MLLM structures, thereby validating the effectiveness of our approach.

**Scale Compression Mechanism.** We compared the proposed Scale Compression Mechanism with the related work FastV [6]. For different methods, we compress the number of visual tokens by 50%. For our method and FastV, we further conduct an experiment with 90% compression. Following FastV’s paper, we set the K in FastV as 2. FastV is a plug-and-play method. Therefore, we conducted this experiment without training the model. As illustrated in Tab. 6, when using 50% compression and 90% compression, our method outperformed FastV by 21.5% and 4.4%, respectively, demonstrating its effectiveness. Both our method and FastV are parameter-free. FastV compresses input tokens, including both visual and textual tokens, within Transformer blocks. In contrast, our method is

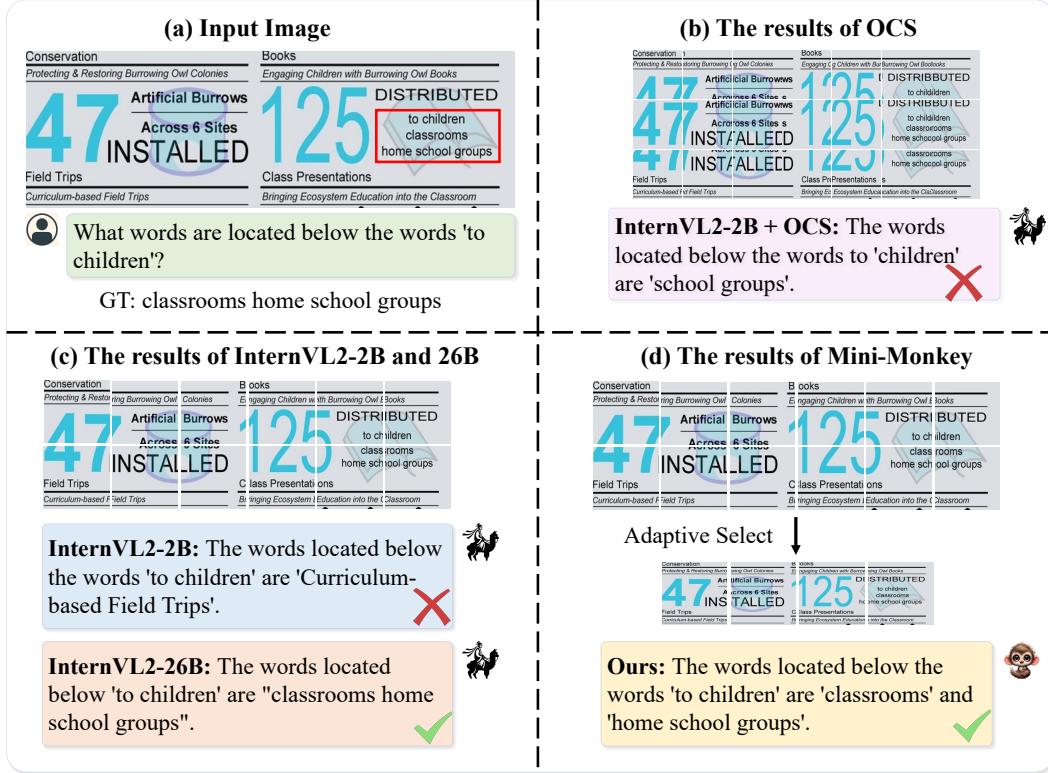


Figure 3: Qualitative results of Mini-Monkey. (a) Input Image and Ground Truth. (b) The results of using overlapping cropping strategy. OSC represents overlapping cropping strategy. (c) The results of InternVL2-2B and InternVL2-26B. (d) The results of Mini-Monkey.

more targeted by using tokens with high relative information density to compress tokens with low information density.

#### 4.4 Qualitative Results

In this section, we provide some qualitative results to demonstrate the effectiveness of our method. First, we verify that the sawtooth effect is particularly evident in lightweight MLLMs, which adopt InternVL2-2B and InternVL2-26B. As shown in Fig. 3(c), InternVL2-26B can answer the questions correctly. However, due to the word ‘classrooms’ and ‘school’ being cropped, InternVL2-2B gives a wrong answer that addresses the text in the bottom left corner of the original image. While Mini-Monkey can overcome this sawtooth effect and provide the correct answer, as presented in Fig. 3(d). Comparing Fig. 3(b) and Fig. 3(d), we can see that the overlapping cropping strategy introduces some hallucinations and cannot answer questions accurately based on the image, whereas our methods can effectively address the sawtooth effect.

## 5 Discussion

There are some other methods to address this issue. One approach is to use a vision encoder that inherently supports high resolution, such as the Swin-Transformer [46] or SAM [30]. However, the cropping strategy remains the most commonly employed method. This preference stems from the ability to leverage the pre-trained, robust vision encoder CLIP [58]. Why high-resolution encoders are not always used to tackle this problem directly? The reason lies in the resource efficiency of CLIP pre-training. Typically, due to the low-resolution input, CLIP requires fewer resources compared to high-resolution visual encoders. Consequently, CLIP is often chosen as the visual encoder in multimodal large language model (MLLM) systems, with the cropping strategy being used to enhance the input resolution.

## 6 Conclusion

In this study, we introduced Mini-Monkey, a lightweight multimodal large language model (MLLM) designed to address the limitations of existing cropping strategies used to enhance MLLMs' ability to process high-resolution images. Traditional cropping methods often segment objects and connected areas, which limits the recognition of small or irregularly shaped objects and text, a problem particularly pronounced in lightweight MLLMs. To mitigate this, Mini-Monkey employs a multi-scale adaptive cropping strategy (MSAC), generating multi-scale representations that allow for the selection of non-segmented objects across different scales. The proposed MSAC can be consistently enhanced across various MLLM architectures. Additionally, we developed a Scale Compression Mechanism (SCM) to reduce the computational overhead of MSAC by compressing image tokens. Our experimental results demonstrate that Mini-Monkey not only achieves leading performance on a variety of general multimodal model understanding tasks but also shows consistent improvements in document understanding tasks. Notably, on the OCRBench benchmark, Mini-Monkey scored 802, surpassing larger 8B-parameter state-of-the-art models like InternVL2-8B. Furthermore, our model and training strategy are exceptionally efficient, requiring only eight RTX 3090 GPUs for training. These results indicate the potential of Mini-Monkey as a powerful and efficient solution for advancing multimodal large language model capabilities in high-resolution image processing.

## References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.
- [2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.
- [3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020.
- [4] Zheng Cai, Maosong Cao, Haojong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.
- [5] Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1511–1520, 2022.
- [6] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. *arXiv preprint arXiv:2403.06764*, 2024.
- [7] Wenhui Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*, 2019.
- [8] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.
- [9] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 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*, pages 24185–24198, 2024.

- [10] Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*, 2023.
- [11] OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. <https://github.com/open-compass/opencompass>, 2023.
- [12] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [13] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model, 2024.
- [14] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, et al. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*, 2024.
- [15] Hao Feng, Qi Liu, Hao Liu, Wengang Zhou, Houqiang Li, and Can Huang. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *arXiv preprint arXiv:2311.11810*, 2023.
- [16] Hao Feng, Zijian Wang, Jingqun Tang, Jinghui Lu, Wengang Zhou, Houqiang Li, and Can Huang. Unidoc: A universal large multimodal model for simultaneous text detection, recognition, spotting and understanding. *arXiv preprint arXiv:2308.11592*, 2023.
- [17] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.
- [18] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.
- [19] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*, 2023.
- [20] Muyang He, Yixin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*, 2024.
- [21] Wenyi Hong, Weihang Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14281–14290, 2024.
- [22] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024.

- [23] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.
- [24] Wenbo Hu, Yifan Xu, Y Li, W Li, Z Chen, and Z Tu. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.
- [25] Runhui Huang, Xinpeng Ding, Chunwei Wang, Jianhua Han, Yulong Liu, Hengshuang Zhao, Hang Xu, Lu Hou, Wei Zhang, and Xiaodan Liang. Hires-llava: Restoring fragmentation input in high-resolution large vision-language models. *arXiv preprint arXiv:2407.08706*, 2024.
- [26] Yulan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091, 2022.
- [27] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018.
- [28] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, pages 235–251, 2016.
- [29] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022.
- [30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.
- [31] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023.
- [32] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.
- [33] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.
- [34] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024.
- [35] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, 2023.
- [36] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26763–26773, 2024.
- [37] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

- [38] Chaohu Liu, Kun Yin, Haoyu Cao, Xinghua Jiang, Xin Li, Yinsong Liu, Deqiang Jiang, Xing Sun, and Linli Xu. Hrvda: High-resolution visual document assistant. *arXiv preprint arXiv:2404.06918*, 2024.
- [39] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.
- [40] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.
- [41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023.
- [42] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- [43] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.
- [44] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023.
- [45] Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024.
- [46] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- [47] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.
- [48] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.
- [49] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.
- [50] Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15630–15640, 2024.
- [51] Pengyuan Lyu, Yulin Li, Hao Zhou, Weihong Ma, Xingyu Wan, Qunyi Xie, Liang Wu, Chengquan Zhang, Kun Yao, Errui Ding, et al. Structextv3: An efficient vision-language model for text-rich image perception, comprehension, and beyond. *arXiv preprint arXiv:2405.21013*, 2024.
- [52] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, 2022.
- [53] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. Infographicvqa. In *WACV*, pages 2582–2591. IEEE, 2022.
- [54] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022.

- [55] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.
- [56] OpenAI. Gpt-4 technical report, 2023.
- [57] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. In *ACL (1)*, pages 1470–1480. The Association for Computer Linguistics, 2015.
- [58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [59] Zhenwei Shao, Zhou Yu, Jun Yu, Xuecheng Ouyang, Lihao Zheng, Zhenbiao Gai, Mingyang Wang, and Jiajun Ding. Imp: Highly capable large multimodal models for mobile devices. *arXiv preprint arXiv:2405.12107*, 2024.
- [60] Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. When do we not need larger vision models? *arXiv preprint arXiv:2403.13043*, 2024.
- [61] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020.
- [62] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.
- [63] Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. Kleister: key information extraction datasets involving long documents with complex layouts. In *International Conference on Document Analysis and Recognition*, pages 564–579. Springer, 2021.
- [64] Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19254–19264, 2023.
- [65] Llama Team. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [66] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [67] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Coglvm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.
- [68] Yonghui Wang, Wengang Zhou, Hao Feng, Keyi Zhou, and Houqiang Li. Towards improving document understanding: An exploration on text-grounding via mllms. *arXiv preprint arXiv:2311.13194*, 2023.
- [69] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language models. *arXiv preprint arXiv:2312.06109*, 2023.
- [70] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, En Yu, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Small language model meets with reinforced vision vocabulary. *arXiv preprint arXiv:2401.12503*, 2024.
- [71] X.ai. Grok-1.5 vision preview. <https://x.ai/blog/grok-1.5v>, 2024.

- [72] Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. LLaVA-UHD: an lmm perceiving any aspect ratio and high-resolution images. *arXiv preprint arXiv:2403.11703*, 2024.
- [73] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.
- [74] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.
- [75] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.
- [76] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [77] Ya-Qi Yu, Minghui Liao, Jihao Wu, Yongxin Liao, Xiaoyu Zheng, and Wei Zeng. Texthawk: Exploring efficient fine-grained perception of multimodal large language models. *arXiv preprint arXiv:2404.09204*, 2024.
- [78] Zhengqing Yuan, Zhaoxu Li, and Lichao Sun. Tinygpt-v: Efficient multimodal large language model via small backbones. *arXiv preprint arXiv:2312.16862*, 2023.
- [79] Jiaxin Zhang, Wentao Yang, Songxuan Lai, Zecheng Xie, and Lianwen Jin. Dockylin: A large multimodal model for visual document understanding with efficient visual slimming. *arXiv preprint arXiv:2406.19101*, 2024.
- [80] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- [81] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023.
- [82] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- [83] Yichen Zhu, Minjie Zhu, Ning Liu, Zhicai Ou, Xiaofeng Mou, and Jian Tang. LLaVA-Phi: Efficient multi-modal assistant with small language model. *arXiv preprint arXiv:2401.02330*, 2024.