MINI-MONKEY: ALLEVIATING THE SEMANTIC SAWTOOTH EFFECT FOR LIGHTWEIGHT MLLMS VIA COMPLEMENTARY IMAGE PYRAMID

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, scaling images to high resolution has received much attention in multimodal large language models (MLLMs). Most existing practices adopt a sliding-window-style cropping strategy to adapt to resolution increase. Such a cropping strategy, however, can easily cut off objects and connected regions, which introduces semantic discontinuity and therefore impedes MLLMs from recognizing small or irregularly shaped objects or text, leading to a phenomenon we call the semantic sawtooth effect. This effect is particularly evident in lightweight MLLMs. To address this issue, we introduce a Complementary Image Pyramid (CIP), a simple, effective, and plug-and-play solution designed to mitigate semantic discontinuity during high-resolution image processing. In particular, CIP dynamically constructs an image pyramid to provide complementary semantic information for the cropping-based MLLMs, enabling them to richly acquire semantics at all levels. Furthermore, we introduce a Scale Compression Mechanism (SCM) to reduce the additional computational overhead by compressing the redundant visual tokens. Our experiments demonstrate that CIP can consistently enhance the performance across diverse architectures (e.g., MiniCPM-V-2, InternVL2, and LLaVA-OneVision), various model capacity (1B→8B), and different usage configurations (training-free and fine-tuning). Leveraging the proposed CIP and SCM, we introduce a lightweight MLLM, Mini-Monkey, which achieves remarkable performance in both general multimodal understanding and document understanding. On the OCRBench, the 2B-version Mini-Monkey even surpasses the 8B model InternVL2-8B by 12 score. Additionally, training Mini-Monkey is cheap, requiring only eight RTX 3090 GPUs. Code and models will be available.

1 INTRODUCTION

Recently, Large Language Models (LLMs) (Zhang et al., 2022; Brown et al., 2020; Touvron et al., 2023; OpenAI, 2023) have received significant attention for their robust text understanding and generation capabilities. Researchers are actively exploring ways to integrate vision encoders into LLMs to upgrade them to multimodal large language models (MLLMs) (Li et al., 2023b; Liu et al., 2023a; Bai et al., 2023). Some approaches employ a Q-former (Alayrac et al., 2022; Li et al., 2023b), while others (Liu et al., 2024d; Wang et al., 2023a) use linear projection. Despite the promising results, they are constrained to processing low-res images, which limits their ability to execute detailed scene analysis.

To address this limitation, much recent effect aims to enable MLLMs to process high-res images. One straightforward solution is to adopt a visual encoder that can tackle high-res images. However, developing a high-quality visual encoder demands substantial training resources (Bai et al., 2023; Chen et al., 2023). An alternative, more resource-efficient strategy is the non-overlapping cropping (Lin et al., 2023; Liu et al., 2024c; Ye et al., 2023a; Li et al., 2024c; Chen et al., 2024b), which splits a high-res image into a set of low-res sub-images.

While the non-overlapping cropping strategy has shown promising results, it inevitably cuts off objects and connected regions, rendering difficulty for the MLLM in recognizing small or irregularly shaped objects due to semantic discontinuity, particularly in the context of document understanding.

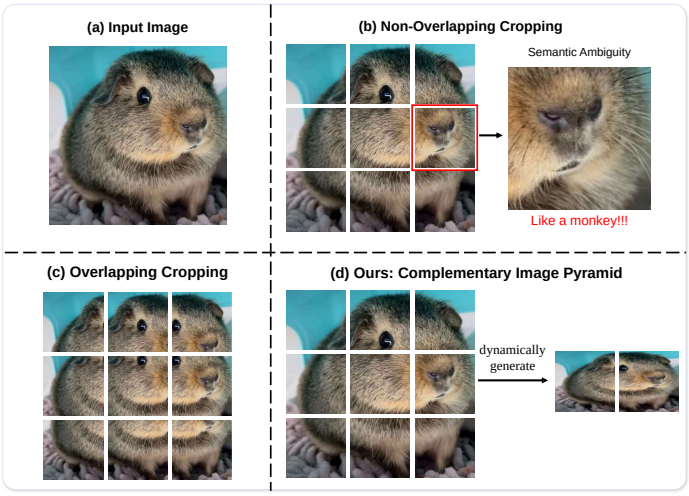


Figure 1: **Comparison of different image cropping strategies**. (a) Input image; (b) Non-overlapping cropping; (c) Overlapping cropping; (d) Ours: complementary image pyramid.

This mainly leads to two types of consequences: i) *semantic ambiguity*: if an object or character is divided, it may be misidentified; the nose of guinea pig in Fig. 1(b) looks much like a monkey after cropping, for instance; 2) *semantic damage*: if a word or sentence is segmented, the meanings of the segmented word will be changed completely; if the word 'breakdown' is divided into 'break' and 'down', the segmented words will have nothing to do with the original one (Liu et al., 2024e; Zhang et al., 2024). For simplicity, we call these phenomena the *semantic sawtooth effect* in this paper. To alleviate this effect, a rather straightforward idea is to adopt overlapping cropping. However, this strategy will result in the processing of much duplicate information, as presented in Fig. 1(c). This redundancy could potentially cause hallucinations in MLLMs. Moreover, it can even deteriorate performance according to our ablation studies in Sec. 4.3. Additionally, the semantic sawtooth effect can be observed more evidently in lightweight MLLMs. Larger MLLMs with enhanced comprehension capabilities and feature extraction capabilities often can alleviate this issue to some extent. Even when the object is segmented, these models can understand the objects through their powerful feature extraction.

To alleviate the semantic sawtooth effect more explicitly, we propose a plug-and-play approach, termed Complementary Image Pyramid (CIP). CIP can be easily integrated into a variety of cropping-based MLLMs, allowing them to tackle high-res images with reduced semantic sawtooth effect. CIP dynamically constructs an image pyramid that provides complementary semantic features for the MLLMs, enabling it rich acquire semantics at all levels. If object semantics are lost at one scale, they can be compensated by those from another scale. Different from previous work (Liu et al., 2024e; Huang et al., 2024) that addresses this issue by modifying the architecture of the model, our approach focuses on enriching the image semantics per se. Consequently, CIP can be easily integrated into a variety of MLLMs, allowing them to tackle high-res images with reduced semantic sawtooth effect. Considering that the CIP introduces some additional computational overheads, we further propose a Scale Compression Mechanism (SCM) for use in situations with limited computational resources. The SCM is both training-free and parameter-free. It leverages the well-trained attention layers of the LLM and the multi-scale information to generate attention weights, which in turn are used to compress redundant tokens. Utilizing the proposed CIP and SCM, we introduce a lightweight MLLM, Mini-Monkey.

Our experiments demonstrate the effectiveness of the proposed method: 1) 2B-parameter Mini-Monkey outperforms the InternVL2-2B by an average of 2.4% across 17 benchmarks in terms of evaluation metrics; 2) Mini-Monkey achieves a score of 806 on the OCRBench, outperforming the 8B-parameter model InternVL2-8B by 12 score. Moreover, we observe that directly fine-tuning well-performing pre-trained MLLM does not enhance, but rather degrades its performance. In contrast, fine-tuning with CIP can facilitate the training process to improve performance. In conclusion, the contributions of this work can be summarized as follows:

- CIP: a plug-and-play complementary image pyramid designed to alleviate the semantic sawtooth effect for multimodal large language models;
- Mini-Monkey: a lightweight, effective, and training-efficient multimodal large language model that integrates the complementary image pyramid and the scale compression mechanism;
- Our method achieves promising results on 8 general multimodal understanding benchmarks and 9 document understanding benchmarks, demonstrating the benefits of alleviating the semantic sawtooth effect.

2 Related Work

2.1 Multimodal Large Language Models

In recent years, Large Language Models (LLMs) have made significant progress (Zhang et al., 2022; Brown et al., 2020; Touvron et al., 2023; OpenAI, 2023). 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(Liu et al., 2024d; Wang et al., 2023a), 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 Q-Former (Li et al., 2023b), Perceiver Resampler (Alayrac et al., 2022), or Abstractor (Ye et al., 2023b), introduce a set of learnable queries to facilitate this integration. Despite these notable advances, previous methods often struggle with detailed scene understanding due to limitations in resolution.

To address this issue, recent research has adopted two primary strategies: 1) The direct use of visual encoders that support high-res input. Some methods (Wei et al., 2023; Hong et al., 2024) utilize two vision encoders, one for processing high-res images and another for low-res images. Others (Wang et al., 2024; Liu et al., 2024f) leverage a single vision encoder. However, these methods require additional training data and parameters, or processing attention over high-res images significantly increases computational demands. 2) An alternative, more resource-efficient method is the cropping strategy, which divides the high-res image into multiple lower-resolution sub-images for processing. Some methods apply a fixed-size cropping scheme (Li et al., 2024c; Lin et al., 2023). Others adopt a dynamic cropping approach (Ye et al., 2023a; Chen et al., 2024b; Dubey et al., 2024) to keep the aspect ratio of the original image. Although the cropping strategy achieves promising results on several multimodal benchmarks, it will inevitably result in a semantic sawtooth effect: 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. This will limit the model's ability to understand the detailed scene. Although some methods (Liu et al., 2024e; Huang et al., 2024) attempt to address this issue by introducing new modules, they introduce additional parameters to the original model and require training this module from scratch. In contrast, the proposed CIP is designed to be seamlessly integrated without introducing additional parameters, offering a plug-and-play solution. A recent work (Shi et al., 2024) also attempts to utilize the multi-scale image to enhance the model's capacity. However, they employ a fixed resolution, lacking the ability to dynamically generate complementary semantics based on the resolution of input images. On the contrary, The CIP dynamically constructs an image pyramid that provides complementary semantic features after cropping for the MLLMs, enabling them to richly acquire semantics at all levels.

2.2 Lightweight Multimodal Large Language Models

Due to the substantial computational costs associated with multimodal large language models (MLLMs), some recent efforts have focused on developing more efficient models for rapid development and real-world applications. For instance, LLaVA-Phi (Zhu et al., 2024) and Imp (Shao et al., 2024) integrate a lightweight large language model with a vision encoder to develop a powerful multimodal system. MobileVLM (Chu et al., 2023) further conserves resources by integrating a lightweight downsampling projector that reduces the number of visual tokens. Bunny (He et al., 2024) advances efficiency through an effective data compression technique, which minimizes the required pretraining dataset. TinyGPT-V (Yuan et al., 2023) adopts a multi-stage training process

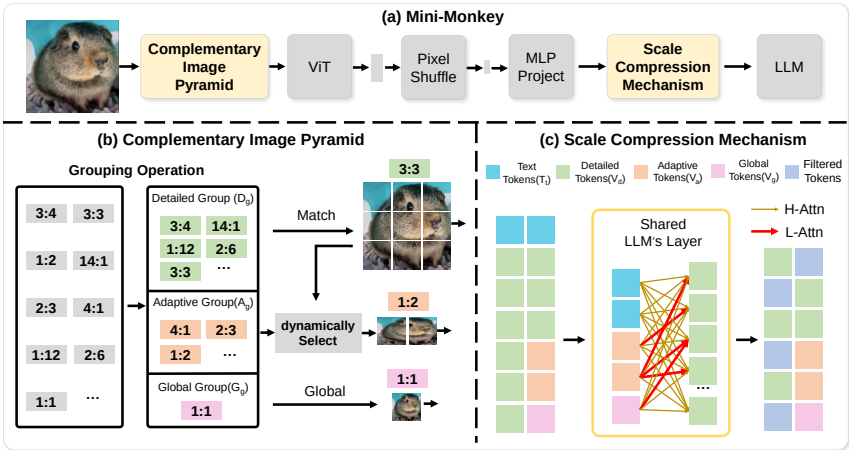


Figure 2: **Overall architecture of Mini-Monkey**. H-Attn represents high attention weights. L-Attn represents low attention weights. The tokens with low attention weights will be filtered. The shared LLM's Layer represents using the block layer from LLM in SCM. The pixel shuffle operation is utilized to reduce the number of visual tokens to one-quarter of the original.

specifically designed for lightweight multimodal models. The above models support only low-res input. To improve detailed scene understanding in lightweight MLLMs, one of the most commonly used methods is the cropping strategy. For instance, MiniCPM-V series (Yao et al., 2024) d employs an adaptive visual encoding method to accommodate high-res images of varying aspect ratios. Similarly, InternVL2-2B (Chen et al., 2024b) enhances the performance of lightweight MLLMs by adopting a dynamic high-res cropping strategy. Despite these advancements, the cropping strategy will introduce a semantic sawtooth effect, which significantly limits the performance of lightweight multimodal large language models. Larger MLLMs with enhanced comprehension capabilities can alleviate this issue to some extent, as discussed in Sec. 4.4.

3 MINI-MONKEY

The overall architecture is illustrated in Fig. 2. Mini-Monkey consists of a CIP, a vision encoder, an MLP layer, a Scale Compression Mechanism, and a Large Language Model (LLM). Initially, CIP dynamically generates an image pyramid based on the resolution of input images. Then, we divide these images into a set of sub-images. These sub-images are then processed by the vision encoder and MLP layer to extract image tokens. The Scale Compression Mechanism adjusts these image tokens and forwards them to the LLM, which subsequently generates the final answers.

3.1 COMPLEMENTARY IMAGE PYRAMID

Existing cropping strategy (Li et al., 2024c; Chen et al., 2024b) directly divides the high-res images into a set of sub-images that will lead to a semantic sawtooth effect. To address this issue, we propose a plug-and-play method, termed complementary image pyramid (CIP), to promote synergy among images at varying scales to alleviate the semantic sawtooth effect. The process of CIP is shown in Fig. 2 (b).

Grouping Operation. We begin by generating a set of pre-defined aspect ratios, which we define as follows: $\{g = (n_h \times n_w) | N_{min} \leq n_h \cdot n_w \leq N_{max}, n_h \in \mathbb{N}, n_w \in \mathbb{N}\}$. The n_h and n_w represent the number of height and width of the grid g. N_{max} is the maximum number of tiles. N_{min} is the minimum number of tiles. These aspect ratios are then categorized into three groups through a grouping operation, including a detailed group D_g, an adaptive group A_g, and a global group G_g. The classification is based on the following criteria: (1) Aspect ratios that are between $\frac{1}{3} * N_{max}$ and N_{max} tiles being allocated to the detailed group, enabling the largest possible image size and thus a clearer depiction of the objects within. (2) For aspect ratios producing between $\frac{1}{8} * N_{max}$ and $\frac{1}{3} * N_{max}$ tiles, we classify them into the adaptive group, which is responsible for enhancing

the fine details at the borders of the crops. (3) The 1:1 aspect ratio is designated to the global group, providing a low-res, comprehensive view of the whole image. The grouping operation generates three groups of different aspect ratios for generating the image.

Dynamically Generating Images. After the grouping operation, we will generate three images from each group. First, we calculate the absolute differences between the aspect ratio of the input image and the aspect ratios within the detailed group. Then, the ratio that has the smallest absolute difference from the input image's aspect ratio is selected as the matched ratio, denoted as D_h, D_w. D_h is the number of height tiles and D_w is the number of width tiles. Once a matched aspect ratio is determined, the image is resized to the corresponding resolution. For example, a 1288×1257 image would be resized to 1344×1344. The resized image is then divided into tiles of 448×448 pixels. After obtaining the detailed image, the adaptive group will dynamically generate an aspect ratio based on D_h, D_w, ensuring that the cropping lines on the detailed group and those on the adaptive group do not overlap. First, we eliminate any aspect ratios in the adaptive group that are exact multiples of D_h, D_w. When the A_h is 1, it means that there is no required cropping. Therefore, they can be integer multiples in such cases. This process can be formulated as follows:

$$\forall k \in \mathbb{Z}, \forall i \in \{h, w\}, \begin{cases} D_i = k \cdot A_i, & \text{if } A_i = 1, \\ D_i \neq k \cdot A_i, & \text{otherwise.} \end{cases} \tag{1}$$

where A_h and A_w denote the height and width components of the aspect ratios in the adaptive group, respectively. Then, we will resize the image by selecting the ratio closest to the aspect ratio of the original image from the remaining aspect ratio.

Because the vision encoder processes each tile independently, existing cropping-based MLLMs fail to capture feature interactions between different tiles. In our method, the adaptive group employs a distinct aspect ratio to partitioning windows compared to the detailed group, thereby simulating cross-tile interaction features and providing the cropping positions information for the detailed group. Similarly, the global group provides the cross-tile interaction features and the cropping positions information for the adaptive component. Three groups provide complementary semantic information and multi-scale information for the model, enabling the model to better capture finer details and handle objects of different sizes in images. Different from the previous method (Liu et al., 2024e; Huang et al., 2024), the proposed CIP alleviates the semantic sawtooth effect from the perspective of the image, bringing several advantages: (i) it is plug-and-play, requiring no additional parameters; (ii) it seamlessly integrates with existing MLLMs that utilize cropping strategies, leading to consistent performance improvements; and (iii) it can be utilized without training and its effectiveness can be further improved through fine-tuning.

3.2 Scale Compression Mechanism

Although the proposed CIP significantly enhances model performance, certain scenarios may restrict the level of computational resources available. 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). The detailed group provides tokens with lower information density, whereas the adaptive and global groups yield tokens that are more information-dense. Therefore, we primarily focus on compressing the tokens from the detailed group. Previous work demonstrates that a well-trained LLM from MLLM can effectively select the necessary visual features based on the input question (Chen et al., 2024a). Consequently, SCM utilizes the layers of the LLM from a well-trained MLLM to select visual tokens without generating any additional parameters. The input visual token including $\mathbf{V_d} \in \mathbb{R}^{L_1 \times C}$, $\mathbf{V_a} \in \mathbb{R}^{L_2 \times C}$, and $\mathbf{V_g} \in \mathbb{R}^{L_3 \times C}$, and the textual token $\mathbf{T_t} \in \mathbb{R}^{T \times C}$ will be sent into an LLM's Layer. $\mathbf{V_d}$ represents the tokens from the detailed group. $\mathbf{V_a}$ represents the tokens from the adaptive group. $\mathbf{V_g}$ represents the tokens from the global group. We utilize the first and second layers of LLM to compress the tokens. The LLM's Layer will output an attention map. We use the visual token from the adaptive group, global group, and textual token to attend to the visual token from the detailed group. The calculation of the attention can be formulated as follows:

$$\mathbf{Q} = \texttt{cat}(\mathbf{V_a}, \mathbf{V_g}, \mathbf{T_t}), \tag{2}$$

$$\mathbf{Attn_w} = \texttt{softmax}(\frac{\mathbf{Q} + \text{PE}(\mathbf{Q}))(\mathbf{V_d} + \text{PE}(\mathbf{V_d}))^T}{\sqrt{D}}). \tag{3}$$

Table 1: Comparison with SoTA models on 8 multimodal benchmarks. General multimodal benchmarks encompass: MME (Fu et al., 2023), RealWorldQA (X.ai, 2024), AI2D test (Kembhavi et al., 2016), CCBench (Liu et al., 2023b), SEED Image (Li et al., 2023a), HallusionBench (Guan et al., 2023), and POPE (Li et al., 2023c). Additionally, the math dataset includes MathVista testmini (Lu et al., 2023). The MME results we report are the sum of the perception and cognition scores. § represents the results from the OpenCompass leaderboard (Contributors, 2023).

model	#param	General Multimodal Benchmarks							Math
		MME	RWQA	AI2D	CCB	SEED	HallB	POPE	MathVista
QWEN-VL (Bai et al., 2023)	7B	1848.3	49.3§	63§	65.7§	52.5§	29.9§	70§	34.9§
Mini-Gemini (Li et al., 2024b)	35B	2141.0	—	—	—	—	—	—	43.3
LLaVA-NeXT (Liu et al., 2024c)	35B	2028.0	—	74.9	49.2	75.9	34.8	**89.6**§	46.5
InternVL 1.2 (Chen et al., 2024c)	40B	2175.4	**67.5**	79.0	59.2	75.6	47.6	88.0	47.7
InternVL 1.5 (Chen et al., 2024b)	26B	**2187.8**	66.0	**80.7**	**69.8**	**76.0**	**49.3**	88.3	**53.5**
DeepSeek-VL (Lu et al., 2024)	1.7B	1531.6	49.7§	51.5§	37.6§	43.7§	27.6§	85.9§	29.4
Mini-Gemini (Li et al., 2024b)	2.2B	1653.0	-	-	-	-	-	-	29.4
Bunny-StableLM-2 (He et al., 2024)	2B	1602.9	-	-	-	58.8	-	85.9	-
MiniCPM-V-2 (Yao et al., 2024)	2.8B	1808.6	55.8§	62.9§	48.0§	-	36.1§	86.3§	38.7
InternVL 2 (Chen et al., 2024b)	2B	1876.8	57.3	74.1	74.7	70.9§	37.9	85.2§	46.3
Mini-Monkey (ours)	2B	**1884.2**	**57.9**	**74.8**	**75.5**	**71.3**	**38.8**	**88.0**	**47.3**

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 first dimension of the attention map $\mathbf{Attn_w} \in \mathbb{R}^{(L_2+L_3+T)\times L_1}$ 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 group, global group, and textual token, are input into the LLM to generate the results. Compared to FastV (Chen et al., 2024a), SCM works in conjunction with the CIP and is more targeted by using tokens with high relative information density to compress tokens with low information density.

4 Experiments

4.1 Implementation Details

We use InternVL2-2B (Chen et al., 2024b) as the Baseline to develop the Mini-Monkey. Following previous work (Chen et al., 2024c), we use the (448, 448) as the input resolution of InternViT. The training datasets used to train the model include DocVQA (Mathew et al., 2021), ChartQA (Masry et al., 2022), DVQA (Kafle et al., 2018), AI2D (Kembhavi et al., 2016), GeoQA+ (Cao & Xiao, 2022), and LLaVA-150K (zh) (Liu et al., 2024d). We use the AdamW (Loshchilov & Hutter, 2017) as the optimizer. The base learning rate is 4e-8. We limit the maximum number N_{max} to 24 and the minimum number N_{min} is 1.

Evaluation. Following the previous work (He et al., 2024; Chen et al., 2024b), we evaluate Mini-Monkey on eleven general multimodal understanding benchmarks, including MathVista testmini (Lu et al., 2023), SEED Image (Li et al., 2023a), RealWorldQA (X.ai, 2024), AI2D test (Kembhavi et al., 2016), POPE (Li et al., 2023c), CCBench (Liu et al., 2023b), MME (Fu et al., 2023), and HallusionBench (Guan et al., 2023). For document understanding, following the previous work (Liu et al., 2024e), we employ two distinct types of metrics to verify the performance of Mini-Monkey. Initially, we leverage the standard metrics provided by the benchmarks to evaluate Mini-Monkey. We utilize benchmarks such as ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), TextVQA (Singh et al., 2019), STVQA (Biten et al., 2019), FUNSD (Jaume et al., 2019), SROIE (Huang et al., 2019), POIE (Kuang et al., 2023) and OCR-Bench (Liu et al., 2023c). We also apply the accuracy metric to verify the performance. Further details on this metric and the used benchmarks can be referenced in appendix A.7.

4.2 Comparison with State-of-the-art Methods

General Multimodal Understanding. We evaluate Mini-Monkey on general multimodal understanding following (He et al., 2024; Chen et al., 2024b). The results are shown in Tab. 1. Mini-Monkey surpasses other 2B-parameter models on 8 benchmarks. These results indicate the ability of Mini-Monkey to handle general multimodal understanding and reasoning tasks. Compared to the

Table 2: Comparison to state-of-the-art MLLMs on OCR-related Tasks. Mini-Monkey achieves the best results among the 2B-parameter MLLMs. § represents the results from the OpenCompass leaderboard (Contributors, 2023).

Model	Model Size	DocVQA^{Test}	ChartQA^{Test}	InfoVQA^{Test}	TextVQA^{Val}	OCRBench
TextMonkey (Liu et al., 2024e)	9B	73.0	66.9	28.6	65.6	558
TextHawk (Yu et al., 2024)	7B	76.4	66.6	50.6	—	—
DocKylin (Zhang et al., 2024)	7B	77.3	46.6	66.8	—	—
HiRes-LLaVA (Huang et al., 2024)	7B	74.7	61.5	48.0	65.4	—
LLaVA-UHD (Xu et al., 2024)	13B	—	—	—	67.7	—
CogAgent (Hong et al., 2024)	17B	81.6	68.4	44.5	76.1	590
UReader (Ye et al., 2023a)	7B	65.4	59.3	42.2	57.6	—
DocOwl 1.5 (Hu et al., 2024a)	8B	82.2	70.2	50.7	68.6	—
HRVDA (Liu et al., 2024a)	7B	72.1	67.6	43.5	—	—
TextSquare (Tang et al., 2024)	7B	84.3	79.4	51.5	66.8	622
IXC2-4KHD (Dong et al., 2024b)	8B	90.0	81.0	68.6	77.2	675
InternVL 1.5 (Chen et al., 2024b)	26B	90.9	**83.8**	72.5	**80.6**	724
InternVL 2 (Chen et al., 2024b)	8B	**91.6**	83.3	**74.8**	77.4	**794**
GLM4-V (GLM et al., 2024)	9B	-	-	-	-	786
Vary-toy (Wei et al., 2024)	1.8B	65.6	59.1	-	-	-
MiniCPM-V 2.0 (Yao et al., 2024)	2.8B	71.9	55.6§	-	74.1	605
InternVL 2 (Chen et al., 2024b)	2B	86.9	76.2	58.9	73.4	784
Mini-Monkey (Ours)	2B	**87.4**	**76.5**	**60.1**	**76.0**	**806**

Table 3: Quantitative accuracy (%) comparison of our model with existing multimodal large language models (MLLMs) on several benchmarks. Following TextMonkey (Liu et al., 2024e), we use the accuracy metrics to evaluate our method.

Method	Scene Text-Centric VQA		Document-Oriented VQA			KIE		
	STVQA	TextVQA	DocVQA	InfoVQA	ChartQA	FUNSD	SROIE	POIE
BLIP2-OPT-6.7B (Li et al., 2023b)	20.9	23.5	3.2	11.3	3.4	0.2	0.1	0.3
mPLUG-Owl (Ye et al., 2023b)	30.5	34.0	7.4	20.0	7.9	0.5	1.7	2.5
InstructBLIP (Dai et al., 2023)	27.4	29.1	4.5	16.4	5.3	0.2	0.6	1.0
LLaVAR (Zhang et al., 2023)	39.2	41.8	12.3	16.5	12.2	0.5	5.2	5.9
BLIVA (Hu et al., 2024b)	32.1	33.3	5.8	23.6	8.7	0.2	0.7	2.1
mPLUG-Owl2-8 (Ye et al., 2024)	49.8	53.9	17.9	18.9	19.4	1.4	3.2	9.9
LLaVA1.5-7B (Liu et al., 2024b)	38.1	38.7	8.5	14.7	9.3	0.2	1.7	2.5
TGDoc (Wang et al., 2023b)	36.3	46.2	9.0	12.8	12.7	1.4	3.0	22.2
UniDoc (Feng et al., 2023b)	35.2	46.2	7.7	14.7	10.9	1.0	2.9	5.1
DocPedia (Feng et al., 2023a)	45.5	60.2	47.1	15.2	46.9	29.9	21.4	39.9
Monkey-8B (Li et al., 2024c)	54.7	64.3	50.1	25.8	54.0	24.1	41.9	19.9
InternVL-8B (Chen et al., 2024c)	62.2	59.8	28.7	23.6	45.6	6.5	26.4	25.9
InternLM-XComposer2-7B (Dong et al., 2024a)	59.6	62.2	39.7	28.6	51.6	15.3	34.2	49.3
TextMonkey-9B (Liu et al., 2024e)	61.8	65.9	64.3	28.2	58.2	32.3	47.0	27.9
InternVL2-2B (Chen et al., 2024b)	65.6	66.2	76.7	46.8	**67.6**	42.0	68.0	66.8
Mini-Monkey-2B (Ours)	**67.2**	**68.8**	**78.4**	**50.0**	67.3	**43.2**	**70.5**	**71.2**

Baseline InternVL2-2B, Mini-Monkey achieves a 1% and 2.8% improvement on the MathVista and POPE, respectively. These results demonstrate that Mini-Monkey not only has superior reasoning skills but also exhibits less frequent hallucinations.

Document Understanding. For the first type of metric, the results are presented in Tab. 2. Compared to Baseline InternVL2-2B, our method outperforms it by 2.6%, 1.2%, and 22 for TextVQA, InfoVQA, and OCRBench, respectively. The CIP provides the model with complementary semantic and multi-scale information, enhancing its ability to perceive fine-grained and varying-sized text. Due to the small original resolution of ChartQA, it is less impacted by cropping operations, resulting in a minor improvement from our method. With the CIP, on the OCRBench, Mini-Monkey even surpasses the 8B-parameter Large Multimodal Model InternVL2-8B and the 9B-parameter Large Multimodal Model GLM4-V by 12 and 20, respectively. These results demonstrate the advantages of CIP in enhancing document understanding. For the accuracy metric, the results are shown in Tab. 3. Mini-Monkey outperforms the InternVL2-2B by 2.6%, 3.2%, and 4.4% on TextVQA, InfoVQA, and POIE, respectively. These results further indicate the great potential of Mini-Monkey for document understanding tasks.

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 (Singh et al., 2019), OCRBench (Liu et al., 2023c), HallusionBench (Guan et al., 2023), MME (Fu et al., 2023), and POPE (Li et al., 2023c) to conduct ablation studies.

Complementary Image Pyramid. We conducted ablation studies to investigate the effectiveness of the CIP. We compared our method with several alternatives: The dynamic high-res strategy (Chen et al., 2024b), which maintains aspect ratios to increase resolution. The fixed-size high-res strategy (Li et al., 2024c), which uses a fixed size to increase resolution. The overlapping cropping

Table 4: Ablation study of Complementary Image Pyramid. We compare our method with the existing cropping strategy and the overlay cropping strategy.

Model	Resolution Strategy	TextVQA	OCRBench	MME	HallB	POPE	Flops(B)	Latency/Example
Baseline	Dynamic high-res Strategy (Chen et al., 2024b)	73.4	784	1876.8	37.9	85.2	**349.4**	**1.0s**
Baseline	Fixed Size high-res Strategy (Li et al., 2024c)	74.2	772	1824.5	37.6	85.0	510.9	1.1s
Baseline	Overlapping Cropping Strategy	70.6	758	1874.1	36.8	83.5	393.1	1.1s
Baseline	Multi-Scale Strategy (Shi et al., 2024)	74.8	776	1846.8	38.1	85.3	559.2	1.6s
Mini-Monkey (Ours)	Complementary Image Pyramid	**76.0**	**806**	**1884.2**	**38.8**	**88.0**	531.3	1.3s

Table 5: Ablation study of the impact of different components in CIP.

	Model	Detailed Component	Adaptive Component	Global Component	TextVQA	OCRBench	MME	HallB	POPE
r1	InternVL2-2B		✓		62.5	385	1686.2	34.8	81.8
r2	InternVL2-2B		✓	✓	70.5	473	1806.1	37.4	86.0
r3	InternVL2-2B	✓			60.8	624	1842.3	37.4	85.3
r4	InternVL2-2B	✓		✓	74.8	782	1874.2	**39.0**	87.5
r5	InternVL2-2B	✓	✓		74.6	785	1853.5	37.6	87.6
r6	InternVL2-2B	✓	✓	✓	**76.0**	**806**	**1884.2**	38.8	**88.0**

Table 6: Ablation study of the different compression rates of SCM.

Compression Rate	0.0	0.1	0.2	0.3	0.4	0.5	0.7	0.9
MME	1884.2	1884.7	1879.8	1878.5	1876.3	1886.0	1871.7	1870.2
Flops (B)	446.9	414.9	383.6	353.0	323.0	293.7	237.0	171.4
Latency/Example	0.83s	0.78s	0.73s	0.67s	0.63s	0.59s	0.51s	0.49s

strategy uses a high-res approach but crops with overlay. The multi-scale strategy (Shi et al., 2024), which introduces a multi-scale strategy to the MLLM. As presented in Tab. 4, the proposed CIP achieved the best results on both general multimodal understanding and document understanding without significantly increasing latency or computational load. The over-overlay cropping strategy, instead of improving the model's performance, actually degraded it.

Various Model Capacity. We performed ablation studies to assess the impact of CIP on models with varying model capacities. As illustrated in Table 8, CIP consistently improves the performance of varying model capacities, illustrating the general applicability of our approach.

Different Usage Configurations. To further validate the improvements introduced by CIP, we performed experiments on usage configurations: a training-free configuration and a fine-tuning configuration. As shown in Tab. 7, CIP demonstrates improvements in performance even when applied without training. The performance can be further improved with fine-tuning. Additionally, we surprisingly find that CIP can even facilitate the model fine-tuning process. As presented in the second line in Tab. 7, direct fine-tuning of the Baseline model not only failed to improve performance but, in some cases, led to a decline. Conversely, incorporating CIP during the fine-tuning of the Baseline resulted in substantial improvements in both general multimodal understanding and document understanding, as evidenced in the fourth line of Tab. 7.

Incorporating CIP to various MLLMs. The proposed CIP can be seamlessly integrated into crop-based MLLMs. To demonstrate its effectiveness, we incorporated CIP into various structures of MLLM, such as MiniCPM-V-2 (Yao et al., 2024), InternVL 2 (Chen et al., 2024b), LLaVA-OV (Li et al., 2024a). As shown in Tab. 8, CIP consistently enhances the performance across different MLLM structures, demonstrating the general applicability of CIP.

The Number of Sub-Images. To investigate whether the performance enhancement is attributed to an increase in the number of sub-images, we performed an experiment by incrementally raising the sub-image count for the Baseline. The findings, summarized in Tab.10, indicate that increasing the number of sub-images does not lead to better performance; instead, it may result in a decline. In contrast, CIP can effectively improve the performance of the model, demonstrating its effectiveness.

The impact of different components in CIP. To evaluate the importance of each component in the CIP, we performed ablation studies using the InternVL2-2B. As presented in Tab. 5, the results indicate that using only the global component or only the detailed component results in a significant performance drop, as shown in r1 and r3 of Tab. 5. By comparing the r1 and r2, as well as r3 and r4, in Tab. 5, we find that adding an Adaptive component significantly improves the performance. When using both the detailed component and the global component, adding the adaptive component leads to further improvements, as shown in r5 and r6 of Tab. 5. Furthermore, the removal of any one of the

Table 7: Exploring different usage configurations of CIP. Train represents fine-tuning the model.

Model	CIP	Train	TextVQA	OCRBench	MME	HallB	POPE
InternVL2-2B	×	×	73.4	784	1876.8	37.9	85.2
InternVL2-2B	×	✓	73.3 (-0.1)	787 (+3)	1858.3 (-18.5)	37.3 (-0.6)	85.3 (+0.1)
InternVL2-2B	✓	×	75.2 (+1.8)	800 (+17)	1881.9 (+5.1)	38.7 (+0.8)	86.7 (+1.5)
InternVL2-2B	✓	✓	76.0 (+2.6)	806 (+22)	1884.2 (+7.4)	38.8 (+0.9)	88.0 (+2.8)

Table 8: Ablation study of incorporating complementary image pyramid (CIP) to other MLLMs. § represents the results from the OpenCompass leaderboard (Contributors, 2023).

Model	CIP	TextVQA	OCRBench	MME	HallB	POPE
MiniCPM-V-2-2.8B	×	74.1	605	1808.6	36.1§	86.3§
MiniCPM-V-2-2.8B	✓	76.0 (+1.9)	627 (+22)	1819.5 (+10.9)	36.5 (+0.4)	87.1 (+0.8)
LLaVA-OV-0.5B	×	65.3	577	1478.0	28.1	86.7
LLaVA-OV-0.5B	✓	66.2 (+0.9)	600 (+23)	1482.6 (+4.6)	28.8 (+0.7)	87.7 (+1.0)
InternVL2-1B	×	70.5	754	1794.4	33.4	84.9§
InternVL2-1B	✓	72.3 (+1.8)	772 (+18)	1801.5 (+7.1)	34.3 (+0.9)	85.7 (+0.8)
InternVL2-2B	×	73.4	784	1876.8	37.9	85.2
InternVL2-2B	✓	76.0 (+2.6)	806 (+22)	1884.2 (+7.4)	38.8 (+0.9)	88.0 (+2.8)
InternVL2-8B	×	77.4	794	2210.3	45.0§	84.2§
InternVL2-8B	✓	79.3 (+1.9)	809 (+15)	2226.4 (+16.1)	45.4 (+0.4)	84.8 (+0.6)

Table 9: Ablation study of the scale compression mechanism. We used different compression ratios to compare with FastV (Chen et al., 2024a). (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 (Chen et al., 2024a) (0.5)	73.4	781	1848.0	38.3	83.9
Mini-Monkey	FastV (Chen et al., 2024a) (0.9)	73.9	792	1866.1	37.5	85.8
Mini-Monkey	SCM (0.5)	74.7	794	**1886.0**	**38.7**	86.1
Mini-Monkey	SCM (0.9)	**75.2**	**801**	1884.7	38.6	**86.2**

three components leads to a decline in performance (r2, r4, r5, and r6). The removal of the global component results in the most significant performance drop (r2 and r6). This is because InternVL2-2B was pretrained with both the detailed and global components. Removing the global component or detailed component will result in a significant performance drop. The adaptive component can to some extent compensate for the information provided by the detailed group, thus the impact of removing the detailed component is less significant than the global component. However, to achieve optimal performance, the synergy among all three components is indispensable. These results demonstrate the effectiveness of our method.

Scale Compression Mechanism (SCM). We compared SCM with the related work FastV (Chen et al., 2024a). 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. As illustrated in Tab. 9, when using 50% compression and 90% compression, our method outperformed FastV by 21.5% and 4.4%, respectively, demonstrating its effectiveness. FastV compresses input tokens, including both visual and textual tokens, within Transformer blocks. In contrast, our method works in conjunction with the CIP and more target by using tokens with high relative information density to compress tokens with low information density.

The different compression rates of SCM. We conduct an ablation experiment on the MME to show how the compression rates of SCM influence the acceleration and computational cost. Following FastV (Chen et al., 2024a), we consider the computation of multi-head attention (MHA) and feed-forward network (FFN) modules in the FLOPs estimation. The total FLOPs are estimated by $L * (4 * n * d^2 + 2 * n^2 * d + 2 * n * d * m)$ where n is the token number, d is the hidden state size, m is the intermediate size of FFN, L is the number of transformer layer. The latency experiments are conducted on A6000 GPU. As presented in Tab. 6, we can find that as the compression ratio increases, the computational load continues to decrease, and the speed keeps improving, without a significant drop in performance. More ablation studies about the selection of hyperparameters are presented in appendix A.

Table 10: Ablation study on the number of sub-images. The number denotes the sub-image count.

Model	Number	TextVQA	OCRBench	MME	HallB	POPE
Baseline	18	74.2	782	1851.7	37.0	85.8
Baseline	24	74.4	783	1857.6	36.9	85.8
Baseline	32	74.3	782	1845.0	36.9	85.9
Baseline	48	74.0	767	1841.6	36.2	85.7
CIP	32	**76.0**	**806**	**1884.2**	**38.8**	**88.0**

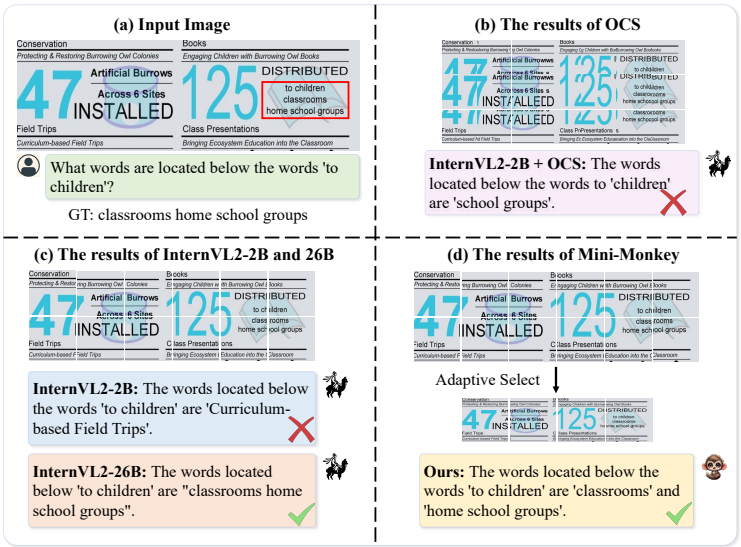


Figure 3: Qualitative results of Mini-Monkey. (a) Input Image and Ground Truth. (b) The results of using overlapping cropping strategy. OSC represents the overlapping cropping strategy. (c) The results of InternVL2-2B and InternVL2-26B. (d) The results of Mini-Monkey.

4.4 Qualitative Results

In this section, we provide some qualitative results to demonstrate the effectiveness of our method. First, we verify that the semantic 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. Mini-Monkey can overcome this semantic 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 semantic sawtooth effect. More qualitative results are presented in appendix A.6.

5 Conclusion

In this study, we introduce a Complementary Image Pyramid (CIP) designed to alleviate the semantic sawtooth effect for MLLMs, thereby enhancing their capability to process high-resolution images effectively. CIP is plug-and-play and can be seamlessly integrated into various multimodal large language models at a low cost. We demonstrate the effectiveness of the proposed CIP across diverse architectures, various parameters, and different usage configurations, leading to consistent performance improvements. Besides, we present a Scale Compression Mechanism (SCM) to compress the visual tokens for computational efficiency. CIP not only enhances the general multimodal understanding performance but also shows consistent improvements in document understanding tasks. Furthermore, our experimental results demonstrate that 2B-parameter MLLM equipped with CIP even surpasses larger 8B-parameter state-of-the-art models like InternVL2-8B on the OCRBench.

REFERENCES

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.

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.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4291–4301, 2019.

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.

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*, pp. 1511–1520, 2022.

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*, 2024a.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023.

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*, 2024b.

Zhe Chen, Jiannan Wu, Wenhai 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*, pp. 24185–24198, 2024c.

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.

OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`, 2023.

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.

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, 2024a.

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*, 2024b.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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*, 2023a.

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*, 2023b.

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.

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.

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.

Muyang He, Yexin 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.

Wenyi Hong, Weihan 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*, pp. 14281–14290, 2024.

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*, 2024a.

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*, 2024b.

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.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. ICDAR 2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1516–1520, 2019. doi: 10.1109/ICDAR.2019.00244.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pp. 1–6, 2019. doi: 10.1109/ICDARW.2019.10029.

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*, pp. 5648–5656, 2018.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, pp. 235–251, 2016.

Jianfeng Kuang, Wei Hua, Dingkang Liang, Mingkun Yang, Deqiang Jiang, Bo Ren, and Xiang Bai. Visual information extraction in the wild: practical dataset and end-to-end solution. In *International Conference on Document Analysis and Recognition*, pp. 36–53. Springer, 2023.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

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*, 2023a.

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*, 2023b.

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*, 2024b.

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*, pp. 292–305, 2023c.

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*, pp. 26763–26773, 2024c.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023.

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*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024b.

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 2024c. URL `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024d.

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*, 2023b.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023c.

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*, 2024e.

Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024f.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.

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.

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.

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*, pp. 2263–2279, 2022.

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*, pp. 2200–2209, 2021.

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*, pp. 1697–1706, 2022.

OpenAI. Gpt-4 technical report, 2023.

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.

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.

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*, pp. 8317–8326, 2019.

Jingqun Tang, Chunhui Lin, Zhen Zhao, Shu Wei, Binghong Wu, Qi Liu, Hao Feng, Yang Li, Siqi Wang, Lei Liao, et al. Textsquare: Scaling up text-centric visual instruction tuning. *arXiv preprint arXiv:2404.12803*, 2024.

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.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023a.

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*, 2023b.

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.

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.

X.ai. Grok-1.5 vision preview. `https://x.ai/blog/grok-1.5v`, 2024.

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.

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.

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*, 2023a.

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*, 2023b.

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.

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.

Zhengqing Yuan, Zhaoxu Li, and Lichao Sun. Tinygpt-v: Efficient multimodal large language model via small backbones. *arXiv preprint arXiv:2312.16862*, 2023.

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.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui 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.

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.

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.

A Appendix

A.1 Ablation study of the maximum number of tiles.

We conduct an experiment to explore the effect of varying the number of tiles in CIP. The results are presented in Table 11. Our findings indicate that overall performance initially improves with an increase in the number of tiles but begins to decline after reaching a certain point. Optimal performance is achieved when the maximum number of tiles is set to 24. Therefore, we choose 24 as the default maximum number of tiles. Additionally, we perform K-means clustering on the resolution ratios of images and use the clustering results as the predefined aspect ratios. We found that using the clustering results as predefined aspect ratios provides a slight improvement over manually preset them.

Table 11: Ablation study of the maximum number of tiles.

maximum number of tiles	TextVQA	OCRBench	MME	HallB	POPE
48	75.4	782	1837.2	39.0	87.5
36	75.7	784	1814.5	**39.1**	87.3
24	76.0	**806**	1884.2	38.8	88.0
12	75.5	796	1874.1	38.8	87.4
6	74.1	788	1879.2	37.9	87.2
K-means	**76.2**	806	**1891.5**	39.1	**88.1**

Table 12: Ablation study of the different settings of grouping operation.

Detailed Group	Adaptive Group	TextVQA	OCRBench	MME	HallB	POPE
$\frac{1}{2} < i < 1$	$\frac{1}{4} < i < \frac{1}{2}$	76.0	800	1886.7	38.7	87.7
$\frac{1}{3} < i < 1$	$\frac{1}{4} < i < \frac{1}{3}$	**76.1**	804	1882.0	38.1	87.8
$\frac{1}{3} < i < 1$	$\frac{1}{8} < i < \frac{1}{3}$	76.0	**806**	**1884.2**	**38.8**	**88.0**
$\frac{1}{4} < i < 1$	$\frac{1}{8} < i < \frac{1}{4}$	75.7	801	1873.7	38.0	87.9
$\frac{3}{4} < i < 1$	$\frac{1}{8} < i < \frac{3}{4}$	75.6	798	1860.6	38.6	87.3

Table 13: Ablation study on the impact of different numbers of LLM layers in SCM.

The number of LLM layers	TextVQA	OCRBench	MME	HallB	POPE	Flops (B)	Latency/Example
6	**75.3**	**798**	**1890.8**	37.6	**88.1**	**489.7**	1.1s
4	75.0	795	1881.2	38.6	86.1	457.0	0.99s
2	74.7	794	1886.0	**38.7**	86.1	424.4	0.92s
1	74.5	789	1878.2	38.3	86.0	408.6	**0.89s**

A.2 Ablation study of different settings of grouping operation.

We conduct experiments to investigate different pre-defined aspect ratio settings for the CIP. All experiments are performed using 24 as the maximum number of tiles N_{tile}. The pre-defined aspect ratios are determined according to the following formula:

$$\{g = (n_h \times n_w) | N_{min} \leq n_h \cdot n_w \leq N_{max}, n_h \in \mathbb{N}, n_w \in \mathbb{N}\}.$$

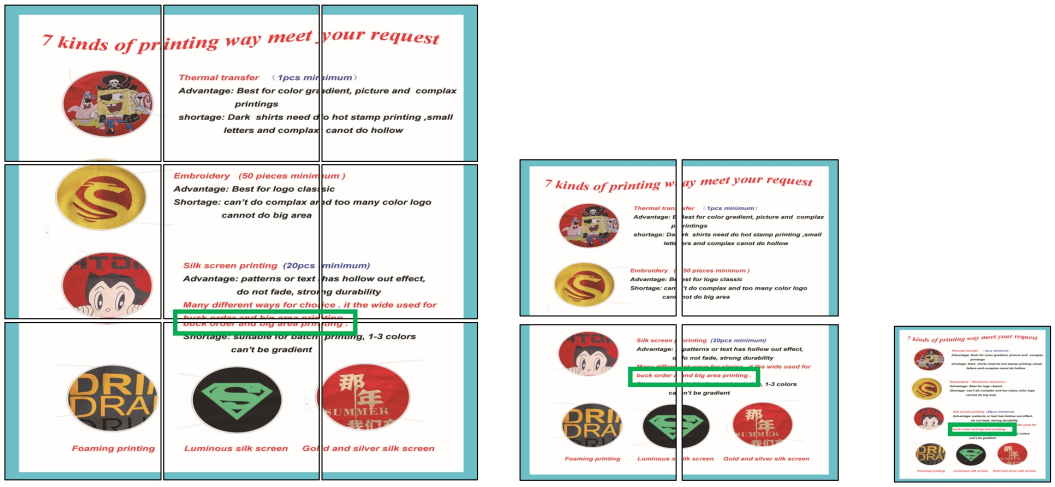
where n_h and n_w represent the height and width of the grid g, respectively. the results are shown in the Tab. 12. $\frac{1}{2} < i < 1$ represents the N_{min} is set to $\frac{1}{2} * N_{tile}$ and the N_{max} is set to $1 * N_{tile}$. According to the results of the experiment, we chose $\frac{1}{3} < i < 1$ for detailed group and $\frac{1}{8} < i < \frac{1}{3}$ for adaptive group. In contrast, the global group employs a fixed 1:1 aspect ratio.

A.3 Ablation study of the different numbers of LLM layers in SCM.

To further investigate the effect of varying the number of layers in LLMs on the compression of visual tokens, we conducted a series of experiments. All experiments are conducted in using 0.5 compression rate. The results are detailed in Tab. 13. Our findings indicate that increasing the number of layers leads to enhanced model performance. Nevertheless, this improvement comes at the cost of increased computational demands and higher latency. Balancing these factors, we decided to adopt a two-layer LLM as our standard configuration, optimizing for both efficiency and performance.

A.4 Effectiveness of Adaptive group

To more intuitively demonstrate the role of the adaptive group, we present a visualization case, as shown in Fig. 4. We can find that the texts are severely corrupted and are hard to read in the detailed image. The global image is used to help retain some of the overall context. However, due to the low-resolution of global images, the texts are blurry. The adaptive group is capable of dynamically adjusting based on the needs of the detailed group and provides more fine-grained feature representations. MLLMs can easily read the texts from the adaptive image.



(a) Detail image: texts have been severely corrupted and are hard to read.

(b) Adaptive image: texts can be recognized easily

(c) Global image: texts are blurry

Figure 4: Qualitative results of CIP. The green box indicates the text that needs to be recognized.

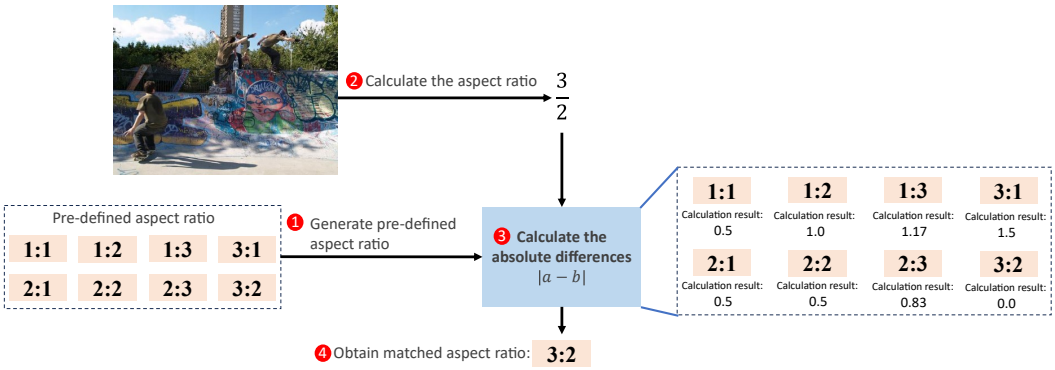


Figure 5: The process of selecting the optimal ratio. First, we generate a set of pre-defined aspect ratios. Next, for a given input image, we calculate its aspect ratio. Then, we calculate the absolute differences between the image's aspect ratio and each of the pre-defined ratios. Lastly, we select the pre-defined aspect ratio that has the smallest difference as the optimal match.

A.5 The Process of Selecting The Optimal Ratio

Given an input image, we first compute the absolute differences between its aspect ratio and those within the detailed group: $|a - b|$. The aspect ratio with the smallest absolute difference to that of the input image is then selected as the matched ratio. The overall process is shown in Fig. 5.

A.6 More Quantitative Results

We present several visualization results from Mini-Monkey, as illustrated in Figure 6. In Figure 6(a), we evaluate the model's performance on general multimodal comprehension. When asked about the characters depicted in the image, Mini-Monkey demonstrated its capability by accurately identifying multiple characters from the Avengers.

In Figure 6(b), we tested the model's understanding of contextual scenarios. By posing questions related to the swimming pool setting, Mini-Monkey not only correctly identified the activities taking place but also provided an insightful analysis of potential hazards associated with the environment, showcasing its ability to infer beyond the visible elements.

Table 14: Quantitative accuracy (%) comparison of our model with existing multimodal large language models (MLLMs) on several benchmarks. Following TextMonkey Liu et al. (2024e), we use the accuracy metrics to evaluate our method.

Method	CIP	Scene Text-Centric VQA		Document-Oriented VQA			KIE			OCRBench
		STVQA	TextVQA	DocVQA	InfoVQA	ChartQA	FUNSD	SROIE	POIE	
MiniCPM-V-2.6-8B		63.8	**73.4**	**82.0**	53.7	69.2	42.9	62.1	80.4	852
MiniCPM-V-2.6-8B	✓	**65.3**	73.1	81.7	**55.3**	**70.6**	**43.4**	**62.5**	**81.1**	**858**

Figure 6(c) highlights Mini-Monkey's proficiency in extracting structured information from images. We tasked the model with converting the visual data into a JSON format, and it successfully produced a detailed and accurate representation, indicating its strong capacity for data organization and structure.

Finally, in Figure 6(d), we assessed the model's ability to process and analyze menu-related information. Mini-Monkey was not only able to precisely recognize and read the text within the image but also effectively understood the context of the questions posed and performed the required mathematical calculations, thereby demonstrating its comprehensive skill set in combining visual and textual analysis.

A.7 The Accuracy Metric.

In this section, we detail the metric described in (Liu et al., 2023c), which establishes a uniform and comprehensible standard for assessing text-focused benchmarks. We utilize benchmarks such as ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), TextVQA (Singh et al., 2019), STVQA (Biten et al., 2019), FUNSD (Jaume et al., 2019), SROIE (Huang et al., 2019), and POIE (Kuang et al., 2023). This metric evaluates outcomes by confirming the presence of the ground truth (GT) within the model's output while excluding any responses that are less than four characters in length. To further validate the effectiveness of our proposed CIP, we have also conducted experiments using this metric with the state-of-the-art method, MiniCPM-V-2.6-8B (Yao et al., 2024). As presented in Tab. 14, our findings reveal that even when applied to a leading multimodal large language model like MiniCPM-V-2.6-8B, the proposed CIP can enhance performance. These results further indicate the effectiveness of the proposed CIP.

A.8 Limitations.

To ensure the seamless application of our method across various architectures, we adopt an image-centric approach to construct an image pyramid without introducing additional parameters. In future work, we will explore the use of trainable Feature Pyramid Network (FPN) for MLLMs, aiming to more efficiently leverage multi-scale features.

Figure 6: Qualitative results of Mini-Monkey. Figures (a) and (b) pertain to general multimodal understanding. Figures (c) and (d) relate to document understanding.