

FALCON: Resolving Visual Redundancy and Fragmentation in High-resolution Multimodal Large Language Models via Visual Registers

Renshan Zhang¹, Rui Shao^{1†}, Gongwei Chen¹, Miao Zhang¹,
 Kaiwen Zhou², Weili Guan¹, Liqiang Nie^{1†}

¹Harbin Institute of Technology, Shenzhen ²Huawei Noah's Ark Lab

zhangrenshan@stu.hit.edu.cn shaorui@hit.edu.cn

chengongwei@hit.edu.cn nieliqiang@gmail.com

<https://github.com/JiuTian-VL/FALCON>

Abstract

The incorporation of high-resolution visual input equips multimodal large language models (MLLMs) with enhanced visual perception capabilities for real-world tasks. However, most existing high-resolution MLLMs rely on a cropping-based approach to process images, which leads to fragmented visual encoding and a sharp increase in redundant tokens. To tackle these issues, we propose the **FALCON** model. **FALCON** introduces a novel **visual register** technique to simultaneously: **1) Eliminate redundant tokens at the stage of visual encoding.** To directly address the visual redundancy present in the output of vision encoder, we propose a Register-based Representation Compacting (**ReCompact**) mechanism. This mechanism introduces a set of learnable visual registers designed to adaptively aggregate essential information while discarding redundancy. It enables the encoder to produce a more compact visual representation with a minimal number of output tokens, thus eliminating the need for an additional compression module. **2) Ensure continuity in visual encoding.** To address the potential encoding errors caused by fragmented visual inputs, we develop a Register Interactive Attention (**ReAtten**) module. This module facilitates effective and efficient information exchange across sub-images by enabling interactions between visual registers. It ensures the continuity of visual semantics throughout the encoding. We conduct comprehensive experiments with **FALCON** on high-resolution benchmarks across a wide range of scenarios. **FALCON** demonstrates superior performance with a remarkable **9-fold** reduction in visual tokens.

1. Introduction

Recently, multimodal large language models (MLLMs) [2, 9, 30, 32, 59, 60] have made significant advancements in

[†]Corresponding authors

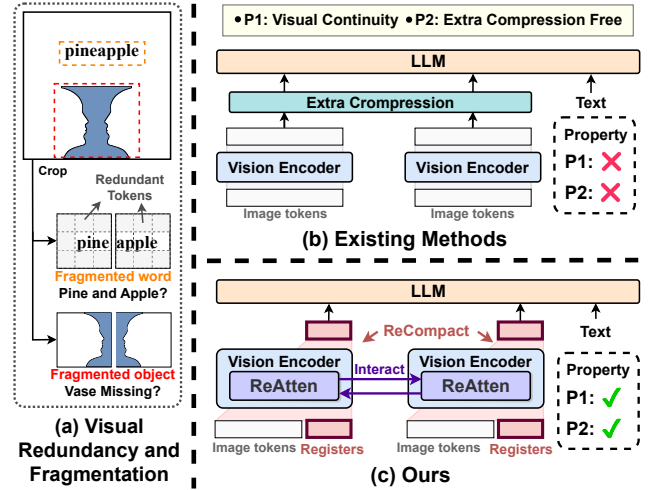


Figure 1. Comparison between existing high-resolution MLLMs and FALCON. FALCON eliminates visual redundancy and ensures the continuity of high-resolution encoding through the use of ReCompact and ReAtten mechanisms.

performing various vision-language tasks [45, 46, 61], facilitating more natural human-machine interactions. Early MLLMs [9, 32, 59] typically processed images at a fixed low resolution, such as 224×224. This limited their capabilities in many real-world applications that require high-resolution input. To equip MLLMs with high-resolution visual perception, recent efforts [15, 17, 33, 58, 67] have introduced a cropping-based high-resolution processing method. Specifically, As shown in Fig. 1(b), this approach crops high-resolution images into multiple non-overlapping sub-images that match the pre-trained resolution of the encoder. Each sub-image is processed independently by the vision encoder. Then the resulting visual features are concatenated and fed into the large language model (LLM).

Although effective, cropping-based methods still face issues of: **1) Visual Redundancy.** The number of visual to-

kens grows dramatically with increasing resolution. While this enriches the visual details, it also leads to an increase in redundant tokens. These redundant tokens significantly increase the computational burden on LLM and interfere with the processing of visual information [65]. To address these issues, as illustrated in Fig. 1(b), existing methods typically rely on additional compression techniques. Representative approaches include combining learnable queries with cross-attention [1, 28, 60] or simply employing pooling layers [48, 55]. However, these methods either demand substantial training resources or fail to fully address redundancy. **2) Visual Fragmentation.** Cropping an image into multiple non-overlapping sub-images may disrupt visual semantics and potentially lead to encoding errors. As shown in Fig. 1(a), in an OCR task, the word “pineapple” may be incorrectly segmented into “pine” and “apple”, causing the LLM to misinterpret the word. Similarly, fragmentation can also hinder the identification of objects. For example, the “rubin vase” pattern illustrated in Fig. 1(a) depicts both a vase and two faces simultaneously. However, the fragmentation could lead to the vase being unrecognizable.

To address these issues, as shown in Fig. 1(c), we propose the **FALCON** model. FALCON introduces a novel **visual register** technique to simultaneously: **1) Eliminate redundant tokens during visual encoding.** To eliminate redundancy directly at the stage of visual encoding, we propose a novel Register-based Representation Compacting (**ReCompact**) mechanism. This mechanism introduces a set of learnable visual registers as additional inputs to the vision encoder. The registers are designed to adaptively aggregate essential information from image tokens and discard redundant content. Then the outputs corresponding to the visual registers are considered as a non-redundant visual representation, serving as the compact input for the LLM instead of a large number of image tokens. This mechanism enables the vision encoder to convey comprehensive visual information in a compact representation using minimal tokens, thus removing the need for an additional compression module. **2) Ensure continuity in visual encoding.** To avoid the potential encoding errors caused by fragmented visual inputs, we develop a novel Register Interactive Attention (**ReAtten**) module. This module leverages the compact and comprehensive representation produced by visual registers to facilitate information exchange across sub-images. It gathers visual registers from different sub-images and employs a specialized cross-ViT attention module to enable interactions among them. Such interaction enables thorough information exchange across the entire image, thus ensuring visual continuity throughout the encoding process. Additionally, the compactness of the visual registers further enhances the efficiency of these interactions. Consequently, this module provides LLMs with a reliable foundation to accurately interpret visual content and reduce hallucinations.

Note that the widely used Abstractor-like modules [28, 33, 58, 60], including QFormer [9], ReSampler [1], follow a similar approach by using query tokens to extract information from image tokens after visual encoding. The key difference is that the visual register mechanism directly adopts the pretrained ViT to aggregate information into a set of specified register tokens. This design is based on the observation that a well-pretrained ViT can naturally gather global information into some specific tokens through self-attention [10]. It offers the visual register mechanism several strengths in terms of: **1) Training Efficiency.** The substantial number of randomly initialized parameters in Abstractor-like modules requires extensive data for pre-training. For instance, QFormer [22], Abstractor [60], and Resampler [1] use large corpora of 129M, 400M, and 1.4B samples, respectively. In contrast, by leveraging well-pretrained ViT parameters, the visual register mechanism requires only 3.5M samples for pretraining. The ablation results in Fig. 4 further demonstrate this advantages. **2) Addressing Visual Fragmentation.** By leveraging the compact and comprehensive representation produced by visual registers, the proposed ReAtten module can effectively facilitate information exchange across sub-images. This is an important function that is notably lacking in Abstractor-like modules. **Combined with ReCompact, it enables the visual register mechanism to simultaneously address the issues of visual redundancy and fragmentation.** The ablation results in Tab. 5 validate the effectiveness of ReAtten.

We summarize our main contributions as follows:

- To address the issue of visual redundancy directly at the encoding phase, we introduce the Register-based Representation Compacting (ReCompact) mechanism. This mechanism enables vision encoder to capture rich visual information using only a minimal number of tokens, effectively eliminating redundancy and obviating the need for additional compression modules.
- To address the issue of visual fragmentation caused by image cropping, we propose the register interactive attention (ReAtten) module. This module ensures the semantic continuity of high-resolution visual encoding in an efficient and effective manner.
- We introduce FALCON, a model that leverages the proposed visual register technique. Comprehensive evaluations on high-resolution visual understanding benchmarks highlight the superiority of FALCON. Notably, FALCON achieves exceptional results with a remarkable reduction in visual tokens by 9×.

2. Related Work

2.1. Multimodal Large Language Models

Research on MLLMs [1, 2, 30, 47, 59, 60] has made significant strides in recent years. By incorporating image in-

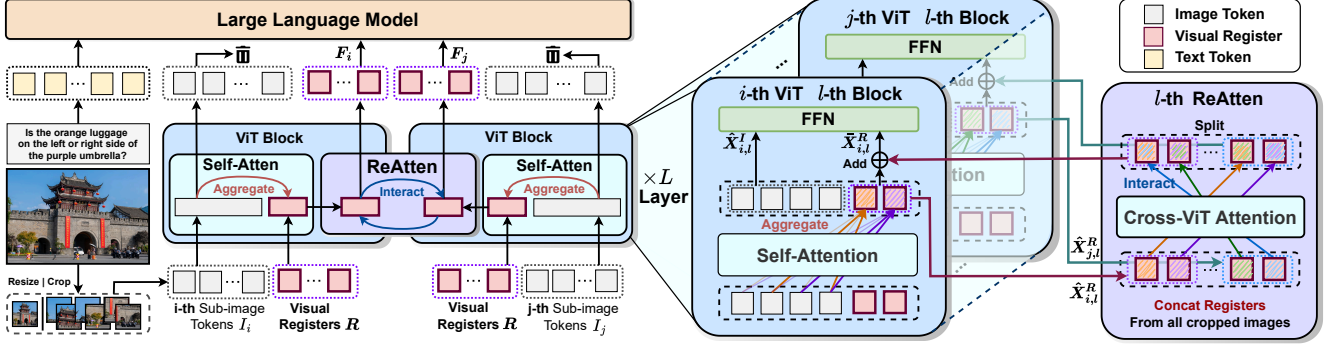


Figure 2. Overall framework of the proposed FALCON. The model introduces a set of learnable visual registers, which are fed into the ViT along with the image tokens from each sub-image. During the visual encoding, the ReCompact mechanism enables information from the image tokens to be aggregated into the visual registers, while the proposed ReAtten facilitates information exchange among the visual registers across different sub-images. Ultimately, these visual registers provide a compact visual representation as input to the LLM.

puts into pretrained LLMs [8, 51, 52], MLLMs demonstrate strong capabilities across various vision-language tasks. Typically, these models use a pretrained vision encoder [43, 62] to transform input images into a set of visual tokens. These tokens are then projected into the input space of the LLM through a projection module (e.g., MLP [30, 32], QFormer [9] and Abstractor [60]) and processed alongside text inputs. Building on the progress of MLLMs, recent studies have shifted from designing specialized methods for certain tasks [44–46] to employing MLLMs as generalist agents, capable of handling diverse practical tasks such as document analysis [15, 57, 64], online video assistant [24], retrieval-augmented generation [40], game agents [27, 29], robot manipulation [18, 21, 39], and GUI agent [3, 5, 54]. Despite these advances, MLLMs still struggle with the fixed low resolution of pretrained image encoders, which can limit their effectiveness in handling high-resolution images.

2.2. High-resolution MLLMs

To address the limitation of MLLMs in handling only low-resolution visual inputs, early studies attempted to either directly incorporate vision encoders that were pre-trained on higher-resolution images [36, 38], or to improve the resolution of pre-trained encoders through additional training [1, 60]. While achieving some progress, they still remain restricted by the fixed resolutions. To accommodate the varying resolutions and aspect ratios, cropping-based techniques have been widely adopted in recent works [15, 17, 33, 58, 67]. However, these approaches still face challenges related to visual redundancy and fragmentation.

To mitigate the issue of redundancy, existing models typically require an additional compression module. For example, Abstractor-like modules utilizing learnable queries combined with cross-attention are widely adopted [1, 28, 33, 58, 60]. These modules typically require pretraining on massive amounts of data, which diminishes their practical efficacy. In contrast, some approaches use pooling

layers [48, 55] or convolutional layers [15, 36]. Though resource-efficient, they often fail to address the issue of redundancy. In this work, we present an initial exploration to directly eliminate redundancy during the visual encoding phase, thereby removing the need for additional compression. To address the issue of fragmentation, several concurrent studies also make preliminary investigation. TextMonkey [33] uses Shifted Window Attention [35] to exchange information between sub-images. But the exchange is still limited to local regions. MiniMonkey [16] proposed a Complementary Image Pyramid strategy to generate complementary cropping schemes. However, this strategy inevitably encodes the same visual content multiple times, further exacerbating visual redundancy. In this work, we propose the ReAtten module to facilitate continuous visual encoding through the interaction among visual registers.

3. FALCON

The overall framework of FALCON is illustrated in Fig. 2. The method begins with shape-adaptive cropping [58], where the image is cropped into multiple non-overlapping sub-images, each matching the resolution of the pre-trained vision encoder. Additionally, a resized version of the full image is used to supplement global layout information. To directly eliminate redundancy during visual encoding, we propose Register-based Representation Compacting (ReCompact) mechanism. This mechanism introduces a set of learnable visual registers, which are paired with image tokens from each sub-image and fed into the vision encoder to capture rich visual information. To ensure the continuity of visual semantics throughout the encoding, a novel Register Interactive Attention (ReAtten) module is integrated into the Vision Transformer (ViT) [11] to facilitate information exchange between sub-images via the visual registers. Finally, the compact visual representations produced by the visual registers are processed through a simple MLP module before being fed into the LLM for further analysis.

3.1. Mitigating Visual Redundancy via ReCompact

Existing MLLMs typically use a ViT-based vision encoder, such as CLIP [43]. For an input image of size $H \times W$, it is patchified into multiple $P \times P$ patches. Each patch is treated as an individual image token, resulting in $N = \frac{H}{P} \times \frac{W}{P}$ tokens, denoted as $I = \{i_1, \dots, i_N\}$. In most current MLLMs, output features corresponding to all image tokens are usually preserved in their entirety. However, when high-resolution images are included, the number of image tokens increases dramatically, placing a substantial burden on the computational resources of LLMs. Consequently, most high-resolution MLLMs necessitate an additional compression module, which can be resource-inefficient. Previous research has shown that most image tokens contain only localized information from their surrounding areas [10]. It suggests that a significant portion of these tokens may be redundant, particularly those originating from background regions. This observation motivates us to directly eliminate redundancy in the output of the vision encoder, thereby reducing the number of visual tokens and avoiding the need for an additional compression module.

To achieve this, we propose a novel Register-based Representation Compacting (**ReCompact**) mechanism. This mechanism employs a set of visual registers to aggregate essential information from image tokens, resulting in a more compact visual representation. Specifically, as shown in Fig. 2, we introduce a set of learnable visual registers, denoted by $R = \{r_1, \dots, r_M\}$, where M represents the number of visual registers, and $M \ll N$. Given a single input image, the image tokens are concatenated with visual registers to create a new input for ViT, represented as $V = \text{Concat}(I, R) = \{i_1, \dots, i_N, r_1, \dots, r_M\}$. During visual encoding, the self-attention mechanism in ViT enables the visual register to adaptively aggregate important information from image tokens while discarding redundant content. Denote $X_{k,l} \in \mathcal{R}^{(N+M) \times D}$ as the input hidden states of the k -th sub-image at the l -th layer, where D is the dimensionality. The self-attention is formulated as:

$$\hat{X}_{k,l} = \text{Self-Atten}(X_{k,l}) = \text{Softmax}\left(\frac{X_{k,l}X_{k,l}^T}{\sqrt{D_{\text{key}}}}\right)X_{k,l} \quad (1)$$

Notably, we do not apply any attention masks for specific constraints. Our experiments show that pretraining on a modest amount of data is sufficient to effectively adapt the pretrained ViT to the proposed ReCompact, enabling it to aggregate essential information within the visual registers. After visual encoding, The output features corresponding to the visual registers, denoted as $F = \{f_1, \dots, f_M\}$, are then used as the visual representation of the image and fed into the LLM for various vision-language tasks.

To further adapt the ReCompact mechanism for cropping-based high-resolution image perception, we share

the same set of visual registers as input for different sub-images. The input for the k -th sub-image is formulated as $V_k = \text{Concat}(I_k, R)$, where $k = 1, 2, \dots, N_c$ and N_c denotes the total number of sub-images. After encoding, the outputs of the visual registers for all sub-images are concatenated, yielding $F_{hr} = \text{Concat}(F_1, F_2, \dots, F_{N_c})$, which serves as the input to the LLM. The ReCompact mechanism enables the vision encoder to represent rich visual information using only a minimal number of M tokens per sub-image, which significantly reduces the visual redundancy and eliminates the need for additional compression.

3.2. Ensuring Visual Continuity via ReAtten

When using the cropping-based high-resolution image encoding method, each sub-image is independently encoded by the vision encoder. This process can disrupt semantic coherence and result in the misrepresentation of visual concepts. To preserve the continuity of visual semantics, it is crucial to enable information exchange between sub-images during encoding. A naive approach would be to directly concatenate the image features from all sub-images and apply self-attention. However, the quadratic computational complexity makes this approach impractical. In Sec. 3.1, we introduced the ReCompact mechanism, where the visual registers effectively aggregate rich visual information while discarding redundancy during visual encoding, allowing them to serve as compact yet comprehensive image representations. Consequently, interactions between visual registers can naturally facilitate sufficient information exchange across different sub-images in an efficient manner.

To enable information exchange between visual registers across different sub-images, we introduce a novel Register Interactive Attention (**ReAtten**) module. As shown in Fig. 2, the ReAtten module is integrated into each of the L layers of the ViT. After the self-attention process in Eq. (1), we acquire $\hat{X}_{k,l} = \{\hat{X}_{k,l}^I, \hat{X}_{k,l}^R\}$, where $\hat{X}^I \in \mathcal{R}^{N \times D}$ and $\hat{X}^R \in \mathcal{R}^{M \times D}$ represent the hidden states corresponding to image tokens and visual registers, respectively. The states of the visual registers from different sub-images are concatenated to form $\hat{X}_l^R = \text{Concat}\left(\{\hat{X}_{k,l}^R\}_{k=1}^{N_c}\right) \in \mathcal{R}^{M \cdot N_c \times D}$. Then these concatenated registers are processed via ReAtten to enable information exchange across sub-images, which can be formulated as:

$$\bar{X}_l^R = \text{ReAtten}(\hat{X}_l^R) = \hat{X}_l^R + \text{Cross-ViT-Atten}(\hat{X}_l^R) \quad (2)$$

where Cross-ViT-Atten is implemented as a self-attention layer. Once the information exchange is completed, the registers and image tokens from a same sub-image are concatenated and then passed into the feedforward network (FFN) for transformation, yielding the output of the l -th layer:

$$X_{k,l+1} = \text{FFN}\left(\text{Concat}(\hat{X}_{k,l}^I, \bar{X}_{k,l}^R)\right) \quad (3)$$

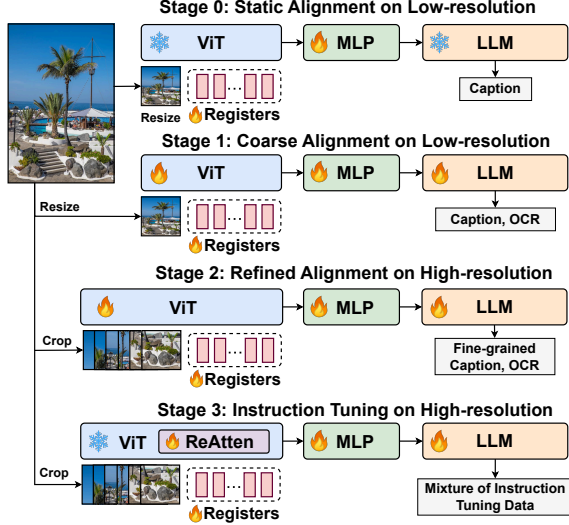


Figure 3. The progressive training pipeline. **Stage 0:** Align the frozen vision and language backbones by training only the MLP and registers. **Stage 1:** Tune all parameters with caption and OCR data on low-resolution for coarse alignment. **Stage 2:** Integrate high-resolution input along with fine-grained captions and OCR to achieve refined alignment. **Stage 3:** Freeze ViT and introduce ReAtten module for high-resolution instruction tuning.

By integrating ReAtten into each layer of the ViT model, FALCON enables effective information exchange between cropped sub-images through visual registers, thereby ensuring semantic continuity of the high-resolution image throughout the visual encoding process.

3.3. Learning Expressive Visual Representation via Progressive Training Pipeline

To empower the ReCompact mechanism with the capability to capture rich visual patterns and to enable smoother initialization of the ReAtten module under high-resolution conditions, we develop a specialized progressive training pipeline. As illustrated in Fig. 3, this pipeline progressively transitions from low to high resolution across four stages.

Stage 0: Static Alignment on Low-resolution. Since both the visual registers and MLP projector are randomly initialized, in stage 0, we aim to achieve a initial alignment between the pretrained vision encoder and the LLM. To facilitate this, we begin by using low-resolution images as inputs, which are directly resized to match the resolution of the vision encoder. We then statically pretrain the model by keeping both the vision encoder and LLM frozen, tuning only the visual registers and MLP using image caption data.

Stage 1: Coarse Alignment on Low-resolution. To better adapt the pre-trained vision encoder to our newly introduced ReCompact mechanism, in stage 1, we unfreeze all model parameters to enable a more thorough tuning. To facilitate the visual registers to progressively learn capturing visual information from coarse to fine granularity, we con-

tinue to use low-resolution inputs and employ high-quality, extended captions rich in visual content, along with full-text OCR data for training. This design allows the registers to acquire holistic scene understanding ability.

Stage 2: Refined Alignment on High-resolution. To further enhance the capacity of the visual register to capture fine-grained visual details, in stage 2, we transition the model to process high-resolution image for refined alignment. Additionally, we employ datasets that includes a series of fine-grained captions and OCR tasks, such as regional captions and text location. This facilitates the visual registers to capture comprehensive fine-grained details.

Stage 3: Instruction Tuning on High-resolution. To adapt the well-aligned visual register for a wide range of real-world tasks, in stage 3, we perform instruction tuning with a mixture of datasets covering various tasks, such as VQA, document understanding, and science reasoning. Additionally, the proposed ReAtten module is integrated to address visual fragmentation caused by cropping-based high-resolution image encoding. Since the vision encoder with ReCompact is well-aligned, the vision encoder is kept frozen and other parameters including ReAtten are trainable during instruction tuning. Notably, to facilitate a smoother initialization for the newly introduced ReAtten modules, we initialize parameters of each ReAtten module with the ViT self-attention parameters at the corresponding depth. This design enables ReAtten to effectively leverage pre-trained parameters, avoiding the need for extensive fine-tuning.

4. Experiments

We implement FALCON with SigLIP-L/16-384px [62] as the vision encoder and Llama-3.1-8B-Instruct [12] as the LLM backbone. A two-layer MLP module with GeLU [13] activation is served as the vision-language projection layer. The number of visual registers is set to 64, achieving remarkable visual token compression rates of $9\times$. The maximum number of sub-images is set to 16.

4.1. Experimental Results

We evaluate FALCON using the recently proposed MME-RealWorld [66] benchmark, a comprehensive and large-scale benchmark specifically designed to assess the high-resolution understanding capabilities of MLLMs. MME-RealWorld comprises 13,366 high-resolution images with an average resolution of $2,000\times 1,500$ pixels, spanning five domains: OCR in the wild, Remote Sensing, Diagrams and Tables, Monitoring, and Autonomous Driving. This benchmark is well-suited for evaluating the practical high-resolution understanding capabilities of MLLMs.

As shown in Tab. 1, FALCON significantly outperform other advanced MLLMs in both perception and reasoning tasks. FALCON achieves state-of-the-art performance with only 64 tokens per sub-image, resulting in a remarkable vi-

Table 1. Evaluation on MME-RealWorld Benchmark. “OCR”, “RS”, “DT”, “MO”, and “AD” stand for OCR in the Wild, Remote Sensing, Diagram and Table, Monitoring, and Autonomous Driving, respectively. “Avg” denotes the weighted average accuracy across all samples, while “Avg-C” denotes the weighted average accuracy across different domains. “Size” indicates the image resolution supported by the model, where $\times n$ signifies a cropping-based high-resolution processing approach with up to n sub-images. “Token” denotes the number of visual tokens after compression. † For models utilizing hybrid vision encoders, “Size” represents the setting of encoder with the highest resolution. We gray out the models that are either closed-source or have at least **twice** the number of visual tokens compared to ours. Among these, the lighter gray denotes scores below ours. The underline indicates that the second performance is achieved. Notably, FALCON achieves state-of-the-art performance with just **64** tokens per sub-image, surpassing most models that have large amount of visual tokens and many advanced closed-source commercial models.

Type	Method	Size ^v	Token ^c	Perception						Reasoning						
	Task Split			OCR	RS	DT	MO	AD	Avg	Avg-C	OCR	DT	MO	AD	Avg	Avg-C
Visual Tokens ≤ 1024	MiniGPT-v2 [4]	448 ²	256	39.0	23.3	20.4	19.3	26.0	26.9	25.6	30.0	20.4	16.9	23.7	23.0	22.7
	Qwen-VL-Chat [1]	448 ²	256	32.4	15.1	15.6	22.1	15.1	20.8	20.1	28.6	13.6	16.5	24.6	22.0	20.8
	LLaVA1.5-13B [30]	336 ²	576	44.1	23.3	20.2	20.5	26.1	28.4	26.8	30.2	20.8	27.5	24.8	25.5	25.8
	ShareGPT4V-13B [6]	336 ²	576	44.6	23.1	20.2	19.3	26.1	28.4	26.6	26.0	20.8	27.3	24.6	24.6	24.7
	TextMonkey [33]	448 ² × 4	256 × 4	37.3	11.7	5.9	16.1	14.3	18.2	17.1	30.4	2.2	4.4	20.0	16.0	14.3
	Mini-Gemini-7B-HD† [26]	336 ² × 4	576	42.0	31.3	22.3	34.2	24.8	31.1	30.9	35.4	24.6	25.9	23.3	26.1	27.3
	DeepSeek-VL† [36]	1024 ²	576	49.6	25.5	23.4	27.0	33.4	33.1	31.8	45.2	23.8	16.7	27.3	28.0	28.3
	MiniCPM-V 2.5 [56]	448 ² × 9	96 × 9	66.8	27.7	52.8	38.7	34.2	47.4	44.0	44.0	31.8	37.0	31.0	34.5	36.0
	Monkey [28]	448 ² × 4	256 × 4	54.6	25.0	32.5	28.0	29.7	36.3	34.0	27.2	20.8	27.3	33.0	28.8	27.1
FALCON	384 ² × 16	64 × 16	<u>66.4</u>	49.6	<u>47.2</u>	40.4	36.5	50.3	48.0	55.4	39.2	44.4	34.5	40.7	43.4	
Visual Tokens ≥ 2048	LLaVA-Next-Qwen72B [31]	336 ² × 4	576 × 4	37.1	29.1	27.7	29.4	18.0	29.0	28.3	17.2	34.2	27.3	29.7	27.9	27.1
	LLaVA-Next-llama3 [31]	336 ² × 4	576 × 4	47.9	25.4	26.6	19.5	18.7	30.1	27.6	55.2	23.4	21.1	30.7	32.1	32.6
	mPLUG-DocOwl 1.5 [15]	448 ² × 9	256 × 9	51.2	23.7	29.3	25.0	28.3	33.7	31.5	42.6	19.8	20.5	26.0	26.9	27.2
	CogVLM2-llama3-Chat [14]	1344 ²	2304	70.0	28.8	47.5	33.7	30.2	45.8	42.0	54.0	32.8	41.2	31.2	37.3	39.8
	IXC2.5 [63]	560 ² × 24	400 × 24	69.3	36.1	63.9	39.5	33.6	52.5	48.5	53.4	41.0	17.7	30.0	33.9	35.5
	LLaVA-OneVision [20]	384 ² × 9	729 × 9	78.7	53.5	60.7	40.3	45.8	59.6	55.8	61.8	40.0	40.8	34.1	41.2	44.2
Closed-source	GPT-4o-mini	-	-	62.5	6.7	44.2	26.5	24.2	37.1	32.8	47.0	39.8	25.8	26.8	32.5	34.9
	Gemini-1.5-pro [50]	-	-	67.6	14.0	39.9	31.1	26.6	39.6	35.9	52.7	33.2	28.3	19.2	29.2	33.4
	GPT-4o	-	-	77.7	28.9	46.7	33.9	22.4	46.4	41.9	61.4	44.8	36.5	26.4	37.6	42.3
	Claude 3.5 Sonnet	-	-	72.5	25.7	67.4	32.2	40.8	52.9	47.7	61.9	61.2	41.8	31.9	44.1	49.2

sual compression rate of 9 times. In addition, FALCON also outperforms most models that have twice as many visual tokens. Specifically, FALCON outperforms CogVLM2-llama3-Chat [14] in average scores on both perception and reasoning task, which uses 2304 tokens per-image. It also surpasses IXC2.5 [63] in average scores on the reasoning tasks, which has a maximum token counts of 9600. Notably, FALCON achieves higher average scores in both perception and reasoning tasks compared to three advanced commercial models, GPT-4o, GPT-4o-mini and Gemini-1.5-pro [50], with it notably outperforming all commercial models in the sub-tasks of remote sensing and monitoring. These results demonstrate that the ReCompact and ReAtten technique proposed by FALCON effectively eliminate visual redundancy and ensure visual continuity, allowing comprehensive and detailed visual information to be conveyed with a highly compact visual representation.

To further assess FALCON’s capabilities in high-resolution understanding, we conduct evaluations on more high-resolution benchmarks: V* [53], DocVQA [42], TextVQA [49] and ChartQA [41]. V* focuses on evaluating

models’ capabilities for fine-grained attribute recognition and spatial reasoning. DocVQA challenges models to comprehend text-dense document images. ChartQA evaluates models’ proficiency in interpreting various types of chart, while TextVQA assesses the ability to recognize small text in real-world scenes. As shown in Tab. 2, FALCON demonstrates superior performance across all these benchmarks.

To evaluate the performance of FALCON across a broader range of domains, we further conduct experiments with diverse popular benchmarks. As shown in Tab. 4, POPE [25], ScienceQA [37], MMBench [34] and SEED-Bench [19] are selected to assess models’ capabilities in hallucination, scientific reasoning, and comprehensive ability. FALCON achieves SOTA performance across all these benchmarks. The results demonstrate that FALCON exhibits superior capabilities across a wide range of tasks.

4.2. Ablation Study

We perform extensive ablation studies to validate the effectiveness of the proposed ReCompact and ReAtten.

Ablation on Training Data. To eliminate the influ-

Table 2. Evaluations on diverse high-resolution Benchmarks. “ V^*_{Avg} ” represents the weighted average score for the two sub-tasks in the V^* benchmark. † denotes OCR inputs are utilized.

Model	V^*_{Avg}	DocVQA	TextVQA	ChartQA
Low-resolution MLLMs				
Qwen-VL [1]	-	65.1	63.8	59.0
InstructBLIP [9]	34.0	4.5	50.7	10.9
LLaVA-v1.5-7B [30]	48.7	8.5	58.2†	17.8
mPLUG-Owl2 [60]	-	17.9	58.2†	22.8
High-resolution MLLMs				
UReader [58]	-	65.4	57.6	59.3
Monkey [28]	-	66.5	65.8	59.0
LlaVA-FlexAttn [23]	54.4	-	48.9	-
S2-Wrapper-7B [48]	55.5	-	61.0	-
LlaVA-Next-Vicuna7B [31]	61.8	-	64.9	54.3
SlIME-7B [67]	-	-	64.4	33.9
FALCON	64.4	74.4	66.5	68.8

Table 3. Ablation results on training data. We train FALCON with two-stage pipeline and exactly the same dataset as LLaVA-v1.5 [30]. The results confirm the superiority of the proposed method when the influence of training data is excluded.

Model	SQA	POPE	MMB	GQA	TextVQA	MM-Vet	SEED-Img
LlaVA-v1.5-7B [30]	66.8	85.9	64.3	62.0	58.2	31.1	65.8
Ours	68.9	87.5	66.0	62.5	61.3	31.4	67.8

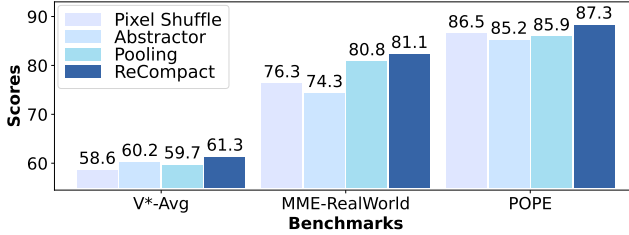


Figure 4. Comparison of different compression methods. We report the sum of the “Avg-C” scores for the perception and reasoning tasks in MME-RealWorld. The superior performance demonstrates that ReCompact effectively eliminates redundancy and produces a more expressive and compact visual representation.

ence of training data and demonstrate the advantages of the proposed method, we utilize exactly the same dataset as LLaVA-v1.5 [30] and employ a two-stage pipeline for training. As shown in Tab. 3, the proposed method exhibits significant performance advantages across multiple benchmarks when trained on the same dataset.

Effectiveness of ReCompact mechanism. To verify the advantages of the proposed ReCompact mechanism in reducing visual redundancy and compressing visual representations during the visual encoding, we compare the ReCompact with other commonly used compression methods. This includes pooling [48, 55], Pixel Shuffle [7] and Abstractor [60], where the Abstractor serves as a representative module that combines query tokens with cross-attention.

As shown in Fig. 4, ReCompact demonstrates superior performance compared to other compression methods.

Table 4. Evaluation on benchmarks across diverse domains. The POPE, SQA, MMBench and SEED-Bench are used to assess hallucination, scientific reasoning, and overall abilities, respectively.

Model	SQA	POPE	MMB	SEED-Img
Low-Resolution MLLMs				
Qwen-VL [1]	67.1	-	41.6	62.3
InstructBLIP [9]	63.1	78.9	38.4	58.8
LlaVA-v1.5-7B [30]	66.8	85.9	62.8	65.8
mPLUG-Owl2 [60]	68.7	85.8	63.5	64.5
High-Resolution MLLMs				
Monkey [28]	69.4	83.7	57.4	64.3
LlaVA-FlexAttn [23]	-	85.9	65.7	-
S2-Wrapper-7B [48]	-	-	66.2	67.9
LlaVA-Next-Vicuna7B [31]	70.1	86.5	66.5	69.6
SlIME-7B [67]	76.8	85.4	63.1	67.0
FALCON	88.8	88.0	76.1	73.3

Table 5. Comparison of different methods for maintaining visual continuity. The “Baseline” model refers to a condition where no specific methods are applied to preserve visual continuity. We report the “Avg-C” score for MME-RealWorld. The results show that the ReAtten module more effectively addresses the issue of visual fragmentation compared to other methods, thereby improving overall performance and reducing hallucinations.

Method	V^*_{Avg}	MME-RealWorld		POPE
		Perception	Reasoning	
Baseline	51.3	38.2	35.6	85.7
CIP [16]	50.3	38.4	37.2	86.3
W-Attn [33]	60.2	41.0	38.1	86.4
ReAtten	61.3	42.1	39.0	87.3

These results indicate that ReCompact effectively eliminates visual redundancy and produces a more expressive compact visual representation. In contrast, we argue that the pooling layer and Pixel Shuffle merely reduces the number of tokens by merging adjacent ones, failing to eliminate redundancy that interferes with the LLM’s understanding of visual content. Additionally, although Abstractor uses more parameters, its performance is generally lower than that of pooling and ReCompact. This result suggests that even millions of samples may be insufficient for fully training Abstractor, which diminishes their practical efficacy.

Effectiveness of ReAtten Module. To validate the advantage of the proposed ReAtten Module, we compare ReAtten with other contemporary methods, including the Shifted Window Attention (W-Attn) [35] adopted by TextMonkey [33] and the Complementary Image Pyramid (CIP) introduced by MiniMonkey [16].

As shown in Tab. 5, the model incorporating ReAtten achieves significant improvements over the baseline across all benchmarks and demonstrates superiority compared to other methods. These results indicate that the ReAtten module effectively enhances the ability of the model in high-resolution perception and mitigates hallucinations by addressing the issue of visual fragmentation. In contrast, we

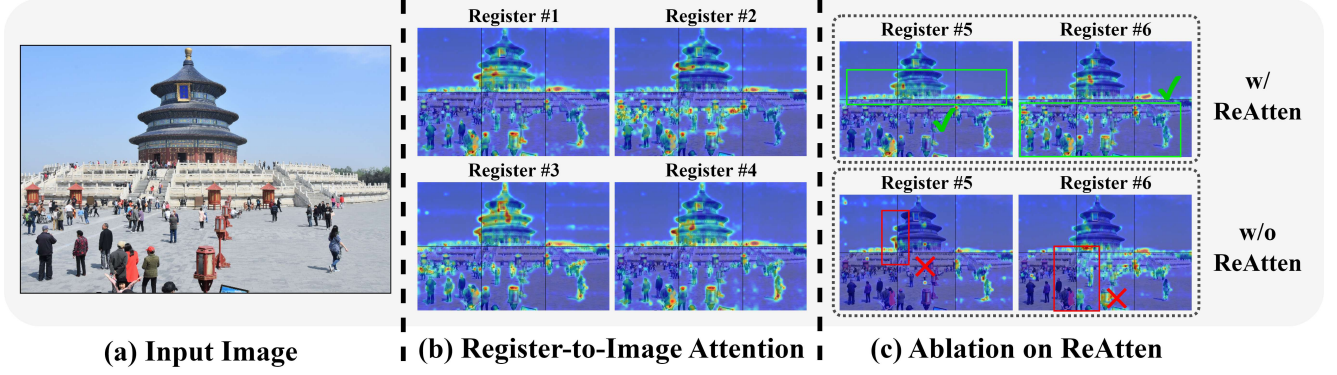


Figure 5. Visualization of the register-to-image attention map. We randomly select visual registers and extract the attention scores that each register assigns to image tokens. These scores are then visualized as heatmaps overlaid on the input image. The results in (b) shows that each visual register focuses on specific parts of the image, capturing rich visual patterns. Meanwhile, registers pay minimal attention to background areas, effectively avoiding the inclusion of redundancy. We also conduct qualitative ablation studies of the ReAtten module. As shown in (c), in the ViT model without ReAtten, the attention patterns across different sub-images appear extremely fragmented. In contrast, the ViT model with ReAtten shows continuous attention patterns, indicating effective information interaction between sub-images.

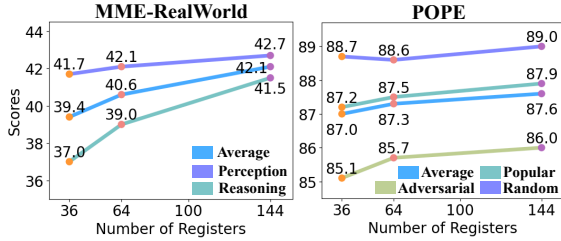


Figure 6. Ablation on different visual register counts. Increasing the number of visual registers generally enhances accuracy across both benchmarks. However, the improvement rate noticeably tapers off when the register count exceeds 64. Based on this observation, to better balance the efficiency and performance, we select register counts of 64 for FALCON.

argue that the information exchange in W-Atten is limited to a local window, which hinders sufficient information exchange among all sub-images. Meanwhile, CIP introduces further redundancy by repeatedly encoding the same visual content, which can even lead to a decline in performance.

Ablation on the Number of Visual Registers. To investigate how visual register count affects model performance, we conduct experiments with visual register counts set to 36, 64, and 144. These counts correspond to visual token compression rates of $16\times$, $9\times$, and $4\times$, respectively. As shown in Fig. 6, increasing the number of visual registers generally enhances model’s high-resolution comprehension capabilities and reduces hallucinations. However, when the register count exceeds 64, the improvement gains start tapering off noticeably. Meanwhile, increasing visual registers demands more computational resources. Compared to the baseline model with 36 registers, models configured with 64 and 144 registers require 1.15 and 2.41 times more training time, respectively. Base on these observations, to achieve a better balance between efficiency and performance, we select 64 as the register counts for FALCON.

4.3. Qualitative Analysis

To qualitatively analyze the proposed visual register technique, we visualize the register-to-image attention maps within ViT. Specifically, for a given visual register, we extract the attention scores it assigns to image tokens in each sub-image and concatenate them to construct a complete heatmap. As shown in Fig. 5(b), each visual register focuses on relatively distinct areas of the image, highlighting that it captures a diverse range of visual details. Moreover, the visual registers tend to avoid background areas, indicating their effectiveness in identifying and discarding visual redundancy. These properties enable visual registers to effectively produce a comprehensive and non-redundant visual representation. Additionally, to qualitatively evaluate the effectiveness of the ReAtten module, we visualize the attention maps within ViT models with and without ReAtten. As shown in Fig. 5(c), in the ViT model without ReAtten, the attention patterns across different sub-images appear highly fragmented. In contrast, the ViT model with ReAtten exhibits continuous attention patterns, indicating effective information exchange between sub-images.

5. Conclusion

To address visual redundancy and fragmentation in high-resolution MLLMs, we propose FALCON. FALCON employs an innovative visual register technique that simultaneously addresses both challenges. This technique uses a Re-Compact mechanism to adaptively aggregate essential visual information through visual registers, creating a compact, non-redundant representation. Additionally, a ReAtten module is introduced to further enhances visual continuity by enabling information exchange among sub-images via visual registers. Extensive experiments demonstrate FALCON’s superiority in high-resolution understanding.

6. Acknowledgement

This study is supported by National Natural Science Foundation of China (Grant No. 62306090, No. U24A20328), Natural Science Foundation of Guangdong Province of China (Grant No. 2024A1515010147), Shenzhen Science and Technology Program (KQTD20240729102207002), Natural Science Foundation of Beijing, China (Grant No. L254018), Jiangsu Science and Technology Major Program (Grant No. BG2024041) and China Postdoctoral Science Foundation (Grant No. 2024M764192).

References

- [1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023. 2, 3, 6, 7
- [2] Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. Lion: Empowering multimodal large language model with dual-level visual knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26540–26550, 2024. 1, 2
- [3] Gongwei Chen, Xurui Zhou, Rui Shao, Yibo Lyu, Kaiwen Zhou, Wentao Li, Yinchuan Li, Zhongang Qi, and Liqiang Nie. Less is more: Empowering gui agent with context-aware simplification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 3
- [4] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 6
- [5] Jingxuan Chen, Derek Yuen, Bin Xie, Yuhao Yang, Gongwei Chen, Zhihao Wu, Li Yixing, Xurui Zhou, Weiwen Liu, Shuai Wang, Kaiwen Zhou, Rui Shao, Liqiang Nie, Yasheng Wang, Jianye HAO, Jun Wang, and Kun Shao. Spa-bench: A comprehensive benchmark for smartphone agent evaluation. In *The Thirteenth International Conference on Learning Representations*, 2025. 3
- [6] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024. 6
- [7] 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*, pages 24185–24198, 2024. 7
- [8] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3):6, 2023. 3
- [9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony 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. 1, 2, 3, 7
- [10] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 4
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3
- [12] 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. 5
- [13] Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *arXiv preprint arXiv:1606.08415*, 2016. 5
- [14] Wenyi Hong, Weihang Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 6
- [15] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3096–3120, 2024. 1, 3, 6
- [16] Mingxin Huang, Yuliang Liu, Dingkan Liang, Lianwen Jin, and Xiang Bai. Mini-monkey: Alleviating the semantic sawtooth effect for lightweight MLLMs via complementary image pyramid. In *The Thirteenth International Conference on Learning Representations*, 2025. 3, 7
- [17] 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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29814–29824, 2025. 1, 3
- [18] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. Openvla: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*, 2024. 3
- [19] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking mul-

- timodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 6
- [20] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. 6
- [21] Hao Li, Qi Lv, Rui Shao, Xiang Deng, Yinchuan Li, Jianye HAO, and Liqiang Nie. STAR: Learning diverse robot skill abstractions through rotation-augmented vector quantization. In *Forty-second International Conference on Machine Learning*, 2025. 3
- [22] 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 *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2
- [23] Junyan Li, Delin Chen, Tianle Cai, Peihao Chen, Yining Hong, Zhenfang Chen, Yikang Shen, and Chuang Gan. Flex-attention for efficient high-resolution vision-language models. In *European Conference on Computer Vision*, pages 286–302, 2024. 7
- [24] Wei Li, Bing Hu, Rui Shao, Leyang Shen, and Liqiang Nie. Lion-fs: Fast & slow video-language thinker as online video assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3240–3251, 2025. 3
- [25] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 6
- [26] 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. 6
- [27] Zaijing Li, Yuquan Xie, Rui Shao, Gongwei Chen, Dongmei Jiang, and Liqiang Nie. Optimus-1: Hybrid multimodal memory empowered agents excel in long-horizon tasks. In *Advances in neural information processing systems*, 2024. 3
- [28] 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. 2, 3, 6, 7
- [29] Zaijing Li, Yuquan Xie, Rui Shao, Gongwei Chen, Dongmei Jiang, and Liqiang Nie. Optimus-2: Multimodal minecraft agent with goal-observation-action conditioned policy. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025. 3
- [30] 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*, pages 26296–26306, 2024. 1, 2, 3, 6, 7
- [31] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 6, 7
- [32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 3
- [33] 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. 1, 2, 3, 6, 7
- [34] 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? In *European Conference on Computer Vision*, pages 216–233. Springer, 2025. 6
- [35] 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. 3, 7
- [36] 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. 3, 6
- [37] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 6
- [38] Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*, 2024. 3
- [39] Qi Lv, Hao Li, Xiang Deng, Rui Shao, Yinchuan Li, Jianye Hao, Longxiang Gao, Michael Yu Wang, and Liqiang Nie. Spatial-temporal graph diffusion policy with kinematic modeling for bimanual robotic manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17394–17404, 2025. 3
- [40] Yibo Lyu, Rui Shao, Gongwei Chen, Yijie Zhu, Weili Guan, and Liqiang Nie. Puma: Layer-pruned language model for efficient unified multimodal retrieval with modality-adaptive learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025. 3
- [41] 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. 6
- [42] 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. 6
- [43] 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. 3, 4

- [44] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10031, 2019. 3
- [45] Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and grounding multi-modal media manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2023. 1
- [46] Rui Shao, Tianxing Wu, Jianlong Wu, Liqiang Nie, and Ziwei Liu. Detecting and grounding multi-modal media manipulation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 3
- [47] Leyang Shen, Gongwei Chen, Rui Shao, Weili Guan, and Liqiang Nie. Mome: Mixture of multimodal experts for generalist multimodal large language models. In *Advances in neural information processing systems*, 2024. 2
- [48] Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. When do we not need larger vision models? In *European Conference on Computer Vision*, pages 444–462, 2024. 2, 3, 7
- [49] 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. 6
- [50] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 6
- [51] 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. 3
- [52] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shriti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 3
- [53] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024. 6
- [54] Bin Xie, Rui Shao, Gongwei Chen, Kaiwen Zhou, Yinchuan Li, Jie Liu, Min Zhang, and Liqiang Nie. Gui-explorer: Autonomous exploration and mining of transition-aware knowledge for gui agent. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. 3
- [55] Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. Deco: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv preprint arXiv:2405.20985*, 2024. 2, 3, 7
- [56] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 6
- [57] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023. 3
- [58] 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 *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2841–2858, 2023. 1, 2, 3, 7
- [59] 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. 1, 2
- [60] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051, 2024. 1, 2, 3, 7
- [61] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision (ECCV)*, pages 69–85. Springer, 2016. 1
- [62] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 3, 5
- [63] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024. 6
- [64] Renshan Zhang, Yibo Lyu, Rui Shao, Gongwei Chen, Weili Guan, and Liqiang Nie. Token-level correlation-guided compression for efficient multimodal document understanding. *arXiv preprint arXiv:2407.14439*, 2024. 3
- [65] Xiaofeng Zhang, Yihao Quan, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. From redundancy to relevance: Enhancing explainability in multimodal large language models. *Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, 2025. 2
- [66] YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, and Rong Jin. MME-realworld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *The Thirteenth International Conference on Learning Representations*, 2025. 5
- [67] Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv preprint arXiv:2406.08487*, 2024. 1, 3, 7