Monkey: Image Resolution and Text Label Are Important Things for Large Multi-modal Models

Zhang Li[1†], Biao Yang[1†], Qiang Liu[2], Zhiyin Ma[1], Shuo Zhang[1], Jingxu Yang[2], Yabo Sun[2], Yuliang Liu[1*], Xiang Bai[1*]
[1]Huazhong University of Science and Technology [2]Kingsoft Office
ylliu@hust.edu.cn

Abstract

Large Multimodal Models (LMMs) have shown promise in vision-language tasks but struggle with high-resolution input and detailed scene understanding. Addressing these challenges, we introduce Monkey to enhance LMM capabilities. Firstly, Monkey processes input images by dividing them into uniform patches, each matching the size (e.g., 448×448) used in the original training of the well-trained vision encoder. Equipped with individual adapter for each patch, Monkey can handle higher resolutions up to 1344×896 pixels, enabling the detailed capture of complex visual information. Secondly, it employs a multi-level description generation method, enriching the context for scene-object associations. This two-part strategy ensures more effective learning from generated data: the higher resolution allows for a more detailed capture of visuals, which in turn enhances the effectiveness of comprehensive descriptions. Extensive ablative results validate the effectiveness of our designs. Additionally, experiments on 18 datasets further demonstrate that Monkey surpasses existing LMMs in many tasks like Image Captioning and various Visual Question Answering formats. Specially, in qualitative tests focused on dense text question answering, Monkey has exhibited encouraging results compared with GPT4V. Code is available at https://github.com/Yuliang-Liu/Monkey.

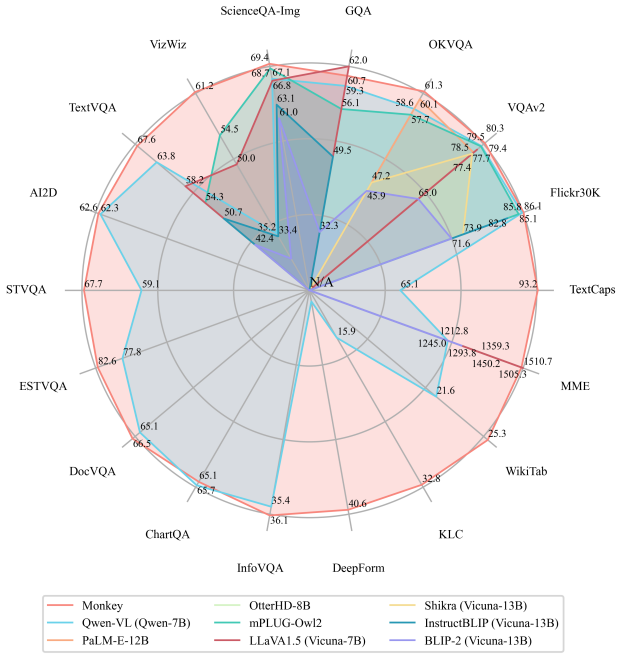

Figure 1. The performance of Monkey on a broad range of multi-modal tasks compared with existing models.

1. Introduction

The field of large multimodal models (LMMs) is advancing quickly because of their skill in handling different types of data, like images and text. Their success in various tasks, including image captioning and visual question answering, is attracting attention in the academic community.

Training LMMs benefits greatly from high-resolution images [3], because higher resolution allows these models to detect more nuanced visual details, leading to accurate recognition of objects, their interrelationships, and the broader context within the image. Additionally, the improved visual clarity of high-resolution images aids in effectively capturing and representing complex details essential for detailed captioning. Despite advancements, handling the wide range of image resolutions and training data quality is still challenging, especially in complex situations. Solutions include using pre-trained visual modules with larger input resolution (like LLaVA1.5 [28]) and gradually increasing the resolution of the training process through

*†equal contribution; *corresponding authors

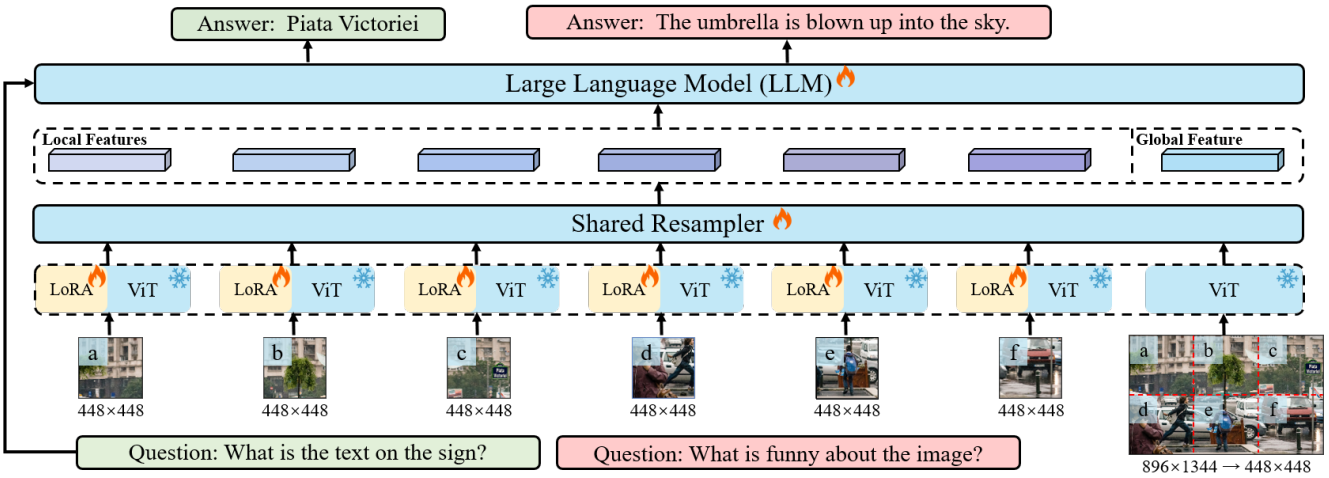

Figure 2. The overall architecture of Monkey. It enables high resolution by capturing global feature from original image and local features from divided patches. All patches are processed through the shared static Vit encoder, such as Vit-BigG with 2b parameters.

curriculum learning (like Qwen-VL [3], PaLI-3 [10] and PaLI-X [9]) have been explored, but they demand significant training resources and still face challenges in handling larger image sizes. To fully leverage the benefits of large input resolution, it is crucial to have more detailed image descriptions, which can enhance the understanding of image-text relationships. However, the short captions in widely used datasets such as COYO [6] and LAION [42] are usually intuitively insufficient.

We introduce Monkey, a resource-efficient approach to increase input resolution within the Large Multimodal Model frameworks. Compared to the approach of directly interpolating the ViT to increase input resolution, Monkey utilizes a new module that divides high-resolution images into smaller patches using a sliding window method. Each patch is processed independently by a static visual encoder, enhanced with LoRA [18] adjustments and a trainable visual resampler. This technique leverages existing LMMs while circumventing the need for extensive pre-training. The key idea is that these encoders are typically trained on smaller resolutions (like 448×448), which is costly to train from scratch. By resizing each patch to its supported resolution, we maintain the training data distribution for the encoder. Our method, which uses various trainable patches to enhance resolution, shows a clear advantage over traditional interpolation techniques for positional embedding, as demonstrated by our quantitative analysis.

To further leverage the advantage of large resolution, we have also proposed an automatic multi-level description generation method. This method is designed to produce high-quality, abundant caption data by seamlessly combining insights from multiple generators. It utilizes the strengths of a diverse array of advanced systems: BLIP2 [27], known for its nuanced image-text understanding; PPOCR [14], a robust optical character recognition system; GRIT [50], which excels in granular image-text alignments; SAM [24], a dynamic model for semantic alignment; and ChatGPT [39], an AI renowned for its contextual understanding and language generation capabilities. By integrating the unique capabilities of these systems, our method offers a comprehensive and layered approach to caption generation, capturing a wide spectrum of visual details.

We summarize the advantages of the Monkey as follows:

1. **Support resolution up to 1344×896 without pretraining**. By going beyond the usual 448×448 resolution used in LMMs, the higher resolution helps to better identify and understand small or closely grouped objects and dense text.
2. **Contextual associations**. We introduce a multi-level description generation method that improves the model's ability to grasp the relationships among multiple targets and more effectively utilize common knowledge in generating text descriptions.
3. **Performance enhancements on many evaluation datasets**. As shown in Fig. 1, we carried out testing across 18 diverse datasets, leading to a very competitive performance by our Monkey model in tasks such as Image Captioning, General Visual Question Answering, Scene Text-centric Visual Question Answering, and Document-oriented Visual Question Answering. In particular, during qualitative evaluations centered on dense text question answering, Monkey has shown promising results, comparing with GPT4V.

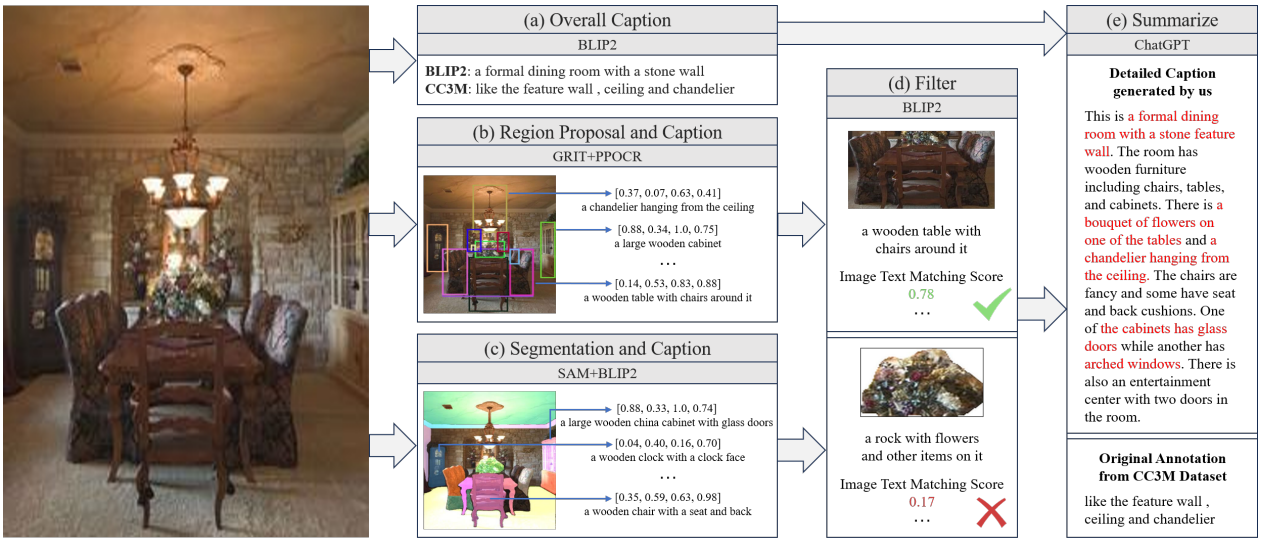

Figure 3. The pipeline for multi-level description generation for images.

2. Related Work

The Large Multimodal Models (LMMs) field has seen significant progress, particularly in enhancing visual and language processing. Methods like Flamingo [1] and OpenFlamingo [2] have advanced visual representation by integrating a Perceiver Resampler with vision encoders. BLIP2 [27] employs a Q-Former to link the frozen LLM and vision encoder. Unified-IO [32] demonstrates versatility by training across over 80 diverse datasets, widening its domain applicability. PaLM-E [13] adopts a unique approach by treating images and text as "multimodal sentences" to improve visual-language tasks. MiniGPT4 [56] bridges visual modules and LLMs, enhancing multimodal capabilities. InstructBLIP [12], starting from BLIP2, adds instructional inputs to the Q-Former for task-relevant visual features. MME [15] introduces a benchmark for evaluating LMMs' perception and cognition.

Additionally, there has been significant progress in leveraging large language models. The LLaVA series, including LLaVA [29] and LLaVA1.5 [28], align vision encoders and LLMs for better image-text understanding. mPLUG-Owl [52] focuses on fine-tuning with mixed text and visual-text data. mPLUG-Owl2 [53] introduces shared modules for better modality collaboration. KOSMOS-2 [41] enables visual answers like detection boxes. Shikra [7] specializes in Referential Dialogue, adept at processing positional inputs and outputs. BLiVA [19] combines task-related and global features for enhanced multimodal task processing. Qwen-VL [3] improves visual module resolution to 448. OtterHD [26] fine-tunes Fuyu-8B [4] with instruction/response pairs, maintaining the original image size during inference.

Despite these advancements, challenges remain in extracting finer image features, as noted by [30, 51], which indicate the need for ongoing development in the field.

3. Methods

Fig. 2 illustrates the comprehensive architecture of Monkey. Initially, the input image is segmented into patches. These patches are then processed through a shared Vision Transformer (ViT) equipped with distinct adapters. Subsequently, both local and global features, along with the question, are processed using the shared resampler and the Large Language Model (LLM), resulting in the generation of the desired answers.

3.1. Enhancing Input Resolution

Input resolution is crucial for accurately interpreting text and detailed image features. Previous studies [3, 10] have shown the effectiveness of starting with smaller resolutions and progressively advancing to larger ones through curriculum learning. However, this approach can be highly resource-demanding, often necessitating comprehensive pretraining with large-scale data (as seen in Qwen-VL, which supports resolutions up to 448×448). To address these issues and efficiently enhance resolution, we introduce a simple yet more effective technique.

Given an image $I \in \mathbb{R}^{H \times W \times 3}$, we employ a sliding window $W \in \mathbb{R}^{H_v \times W_v}$ (where H_v, W_v denote the supported resolution of the original LMM) to partition the image into smaller, local sections. We also leverage LoRA [18] within each shared encoder to address the varied visual elements in different parts of an image. This integration of LoRA is to help our encoders to recognize and assimilate detail-

sensitive features from each image area effectively, which enhances the understanding of spatial and contextual relationships without a substantial increase in parameters or computational demand.

To preserve the overall structural information of input image, we resize the original image to dimensions (H_v, W_v), maintaining it as a global image. Following this, both the individual patches and the global image are processed through the visual encoder and resampler concurrently. Here, the visual resampler, inspired by Flamingo [1], is a mechanism that performs two main functions: summarizing visual information and obtaining higher semantic visual representations in a language feature space. It achieves this by leveraging a cross-attention module. The module employs trainable vectors (embeddings) as query vectors, along with image features from the visual encoder serving as keys for cross-attention operations.

This approach strikes a balance between detailed and holistic perspectives of the images, thereby enhancing the model performance while avoiding a substantial increase in computational demand.

3.2. Multi-level Description Generation

Previous models such as LLaVA [29] and Qwen-VL [3] used large datasets like LAION [42], COYO [6], and CC3M [43] for their initial training. However, these datasets often offer image-text pairs that are too simple (e.g., one short sentence to describe a complicated image), lacking in detailed imagery. As a result, even when these models are trained with high-resolution images, they struggle to accurately link visual features with basic captions. This limitation affects the models to effectively combine visual processing with language understanding.

To bridge this gap, we develop a novel approach for generating multi-level descriptions automatically. This technique is designed to create rich and high-quality caption data by effectively blending the outputs from various generators. We utilize a combination of several advanced systems, each bringing its own strength to the process: BLIP2 [27], which provides a deep understanding of the relationship between images and text; PPOCR [14], a strong performer in optical character recognition; GRIT [50], specializing in detailed image-text matching; SAM [24], focused on semantic alignment; and ChatGPT [39], known for its exceptional ability in contextual language generation.

As shown in Fig. 3, the image description process begins with BLIP2 creating overall captions using a Q-former for tight integration with the vision encoder and LLM, while retaining original CC3M annotations for context. Next, GRIT, a region-to-text model, generates detailed descriptions of specific regions, objects, and their characteristics. PPOCR extracts text from the images, and SAM segments and identifies objects and their parts. These objects are then individually described by BLIP2. However, to counter potential inaccuracies from these tools, especially in zero-shot settings, we find it essential to further use BLIP2 to check for consistency between image areas, objects, and their descriptions, filtering out low-scoring matches. Finally, all data, including global captions, localized descriptions, text extracts, and object details with spatial coordinates, are fed into the ChatGPT API for fine-tuning, enabling ChatGPT to generate accurate and contextually rich image descriptions.

By merging the unique features of these systems, our approach achieves a layered and comprehensive style of caption creation. It captures an extensive range of visual and textual nuances, resulting in captions that are not just elaborate, but also contextually diverse and engaging.

Task	Dataset	Smaples
Image Caption	Detailed Caption	213k
	COCO Caption [22]	82k
	TextCaps [44]	109k
General VQA	VQAV2 [16]	100k
	OKVQA [34]	18k
	GQA [20]	150k
	ScienceQA [33]	18k
	VizWiz [17]	20k
Scene Text-centric VQA	TextVQA [45]	34k
	OCRVQA [38]	250k
	AI2D [23]	24k
Doc-oriented VQA	DocVQA [36]	118k
	ChartQA [35]	84k
	InfoVQA [37]	47k
	DeepForm [47]	7k
	KLC [46]	27k
	WTQ [40]	28k
	TabFact [8]	91k
	VisualMRC [48]	21k
Total	-	1.44m

Table 1. Details on the Monkey training data, derived entirely from publicly available datasets.

3.3. Multi-task Training

Our goal is to train a model that is both cost-effective and capable of understanding different types of images for various tasks. By integrating various datasets and employing uniform instructions for all tasks, as guided by [3], we enhance the model's learning ability and training efficiency.

We focus on tasks such as creating image captions, responding to image-based questions, and other activities requiring the model to process both text and images. For captioning, we instruct the model with "Generate the caption in English:" for basic captions, and "Generate the detailed caption in English:" for more intricate ones. When it comes to answering questions about images, we use a straightfor-

Model	Image Caption		General VQA				
	Flickr30K	TextCaps	VQAv2	OKVQA	GQA	ScienceQA	VizWiz
Flamingo-80B [1]	67.2	-	56.3	50.6	-	-	31.6
Palm-E-12B [13]	-	-	77.7	60.1	-	-	-
BLIP-2 (Vicuna-13B) [27]	71.6	-	65.0	45.9	32.3	61.0	19.6
InstructBLIP (Vicuna-13B) [12]	82.8	-	-	-	49.5	63.1	33.4
Shikra (Vicuna-13B) [7]	73.9	-	77.4	47.2	-	-	-
mPLUG-Owl2 [53]	85.1	-	79.4	57.7	56.1	68.7	54.5
LLaVA1.5 (Vicuna-7B) [28]	-	-	78.5	-	**62.0**	66.8	50.0
Qwen-VL(Qwen-7B) [3]	85.8	65.1	79.5	58.6	59.3	67.1	35.2
Qwen-VL-Chat [3]	81.0	-	78.2	56.6	57.5	68.2	38.9
Monkey	**86.1**	**93.2**	**80.3**	**61.3**	60.7	**69.4**	**61.2**

Table 2. Results on Image Caption and General VQA.

Model	TextVQA	AI2D	STVQA	ESTVQA
Pix2Struct-Large [25]	-	42.1	-	-
BLIP-2 [27]	42.4	-	-	-
InstructBLIP [12]	50.7	-	-	-
mPLUG-DocOwl [52]	52.6	-	-	-
mPLUG-Owl2 [53]	54.3	-	-	-
Qwen-VL [3]	63.8	62.3	59.1	77.8
Qwen-VL-Chat [3]	61.5	57.7	-	-
LLaVA-1.5 [28]	58.2	-	-	-
Monkey	**67.6**	**62.6**	**67.7**	**82.6**

Table 3. Results on Scene Text-centric VQA.

Model	DocVQA	ChartQA	InfoVQA	DeepForm	KLC	WTQ
Qwen-VL	65.1	**65.7**	35.4	4.1	15.9	21.6
Monkey	**66.5**	65.1	**36.1**	**40.6**	**32.8**	**25.3**

Table 4. Results on Doc-oriented VQA.

ward format: "{question} Answer: {answer}."

In our training process, we use a variety of public datasets tailored to specific tasks. For image captioning, we include both our own detailed captions and established datasets like COCO caption [22] and TextCaps [44]. For general Visual Question Answering (VQA), we utilize datasets such as VQAV2 [16], OKVQA [34], GQA [20], ScienceQA [33], and VizWiz [17]. For Text-centric VQA tasks, we select TextVQA [45], OCRVQA [38], and AI2D [23]; while for document-related VQA, we employ datasets like DocVQA [36], ChartQA [35], InfoVQA [37], DeepForm [47], Kleister Charity (KLC) [46], WikiTableQuestions (WTQ) [40], TableFact [8], and VisualMRC [48]. We use our multi-level description generation method to regenerate around 427k image-text pairs from the CC3M dataset, previously used in LLaVA's pretraining phase. To ensure balanced training, we control the image count for each task as detailed in Tab. 1. Our compiled dataset, with around 1.44 million examples, is designed to train our model effectively in understanding and executing various instructions.

4. Experiment

We evaluate our model by testing it across a spectrum of standard vision-language tasks, including the generation of image descriptions, answering diverse visual questions, and comprehending targeted phrases in images.

4.1. Implementation Details

Model Configuration. We conduct experiments based on the well-trained Vit-BigG [21] and LLM from Qwen-VL [3], the pre-trained large multimodal model. Since the vision encoder has already been well pretrained, we proceed directly to the instruction-tuning stage. During instruction tuning, H_v, W_v are set to 448 to match the encoder of Qwen-VL. We employ a consistent resampler across all crops. The learnable queries engage with local features, utilizing the same set of 256 learnable queries for each crop. Due to limitations in training time, our main experiments were mainly conducted using images of size 896×896 unless specify. For LoRA, we set the rank to 16 for the attention module and 32 for MLP in the encoder. Monkey includes 7.7B parameters for a large language model, with 90M parameters for the resampling module, an encoder with 1.9B parameters, and 117M parameters for LoRA. The overall parameters for Monkey is 9.8B.

Training. During the training process, we utilize the AdamW optimizer [31] with a learning rate of 1e-5 and the cosine learning rate schedule. Additionally, we set the values of β_1 and β_2 to 0.9 and 0.95, respectively. We incorporate a warmup period of 100 steps and employ a batch size of 1024. To control overfitting, we apply a weight decay of 0.1. The whole training process takes 40 A800 days for one epoch.

	Resolution	LoRA	Throughout	FLOPS (e20)	VQAv2	GQA	TextVQA	STVQA	DocVQA	DeepForm	InfoVQA	WTQ
r1	896×896*	0	43.452	1.608	74.1	55.2	44.7	41.5	53.9	11.4	32.7	16.8
r2	896×896*	1	37.429	1.614	71.4	54.0	41.7	38.5	47.5	7.2	31.5	17.1
r3	672×672	4	43.604	1.617	80.0	59.6	67.3	67.2	66.4	31.3	35.9	25.0
r4	784×784	4	42.851	1.617	79.9	59.8	67.5	**67.7**	66.5	38.9	35.5	25.1
r5	896×1344	6	28.542	1.622	80.1	61.1	67.3	66.7	66.3	**42.3**	**39.6**	**26.6**
r6	1344×896	6	28.842	1.622	80.2	**61.8**	**67.7**	66.3	64.5	41.4	35.7	25.2
r7	896×896	0	49.634	1.613	80.1	60.4	67.5	65.1	66.1	36.8	36.1	24.9
r8	896×896	1	42.885	1.614	80.0	60.3	67.6	67.0	**66.7**	36.9	36.5	24.7
r9	896×896	4	42.542	1.617	**80.3**	60.7	67.6	**67.7**	66.5	40.6	36.1	25.3

Table 5. Ablation study on enhancing input resolution and the number of trainable adapters using Qwen-VL (originally trained using 448×448). * refers to directly scaling the input size of the visual encoder from 448 to 896 using traditional positional position interpolation.

4.2. Results

We report the results on Image Caption, General VQA, Scene Text-centric VQA, and Document-oriented VQA. We also conduct testing on the MME benchmark and achieve a perception score of 1505.3, ranking second, as shown in Fig. 1. The details of each dataset can be found in Appendix.

Image Caption. Image captioning is vital for connecting visual content with the understanding of natural language. In our study, we select Flickr30K [54] and TextCaps [44] as the benchmark for testing the image captioning task. TextCaps challenges the model to interpret and reason text within images effectively. We present our model's performance on Flickr30K and TextCaps in Tab. 2, where the results indicate that Monkey demonstrates enhanced performance on these datasets. We also qualitatively show effectiveness of our method in offering detailed image descriptions in Sec. 4.4 and Appendix.

General VQA. General visual question answering (VQA) requires ability to learn visual and textual information, showing a deep understanding of how they interrelate. For General VQA, we validate on five benchmarks: VQAv2 [16], OKVQA [34], GQA [20], ScienceQA [33], and VizWiz [17]. The performance results are shown in Tab. 2. Our model shows remarkable proficiency in VQAV2, OKVQA, ScienceQA, and VizViz, surpassing the nearest competing method by an average of 1.62%. These results highlight the effectiveness of our method, emphasizing its use of high input resolution and detailed data.

Scene Text-centric VQA. Text information is commonly found in real-world scenes, making the ability to answer questions about text in images a crucial aspect of question-answering tasks. For our evaluation, we employ four datasets: TextVQA [45], AI2D [23], STVQA [5], and ESTVQA [49]. The results, shown in Tab. 3, indicate that our model leads in performance on these datasets, outperforming the nearest competitor by an average of 4.37%. Based on our observation, this enhanced performance is mainly attributed to the increased image resolution, which brings smaller text and finer details into clearer view. Moreover, the inclusion of detailed caption data during training provides valuable textual context, further boosting the robustness of the model.

Document-oriented VQA. Despite the clean backgrounds of documents, their densely packed text poses distinct challenges. To effectively evaluate our model, we select representative benchmarks including DocVQA [36], ChartQA [35], InfographicVQA [37], DeepForm [47], KLC [46], and WTQ [40]. The results, as detailed in Tab. 4, show that Monkey surpasses Qwen-VL in most Document-oriented VQA tasks, achieving an averagely significant improvement of 9.77%. The higher resolution of documents reveals more intricate details and a denser concentration of information. Monkey's capability to process larger input resolutions enhances its spatial perception, thereby improving its recognition and comprehension of various document elements like text, charts, infographics, and forms.

4.3. Ablation Study

We conduct thorough experiments to validate the effectiveness of our designs.

Ablation study on strategies of enhancing input resolution. We first evaluate the existing technique of improving input resolution, as illustrated in Tab. 5. Resizing the visual encoder using traditional positional position interpolation to a size of 896 results in worse performance compared with our method under the same settings (r1 vs. r9). Interestingly, applying LoRA to the encoder for this traditional interpolation method appears to be less effective than not using it (r1 vs. r2). This may due to the inherited parameters from the previous encoder are specifically tuned by lower resolution, changing it by force may necessitate more training resources.

For our method (r3-r9), as we increase the input size, there is a noticeable boost in performance, especially demonstrated in the DeepForm dataset. It can be observed that adding LORA does not significantly increase FLOPs

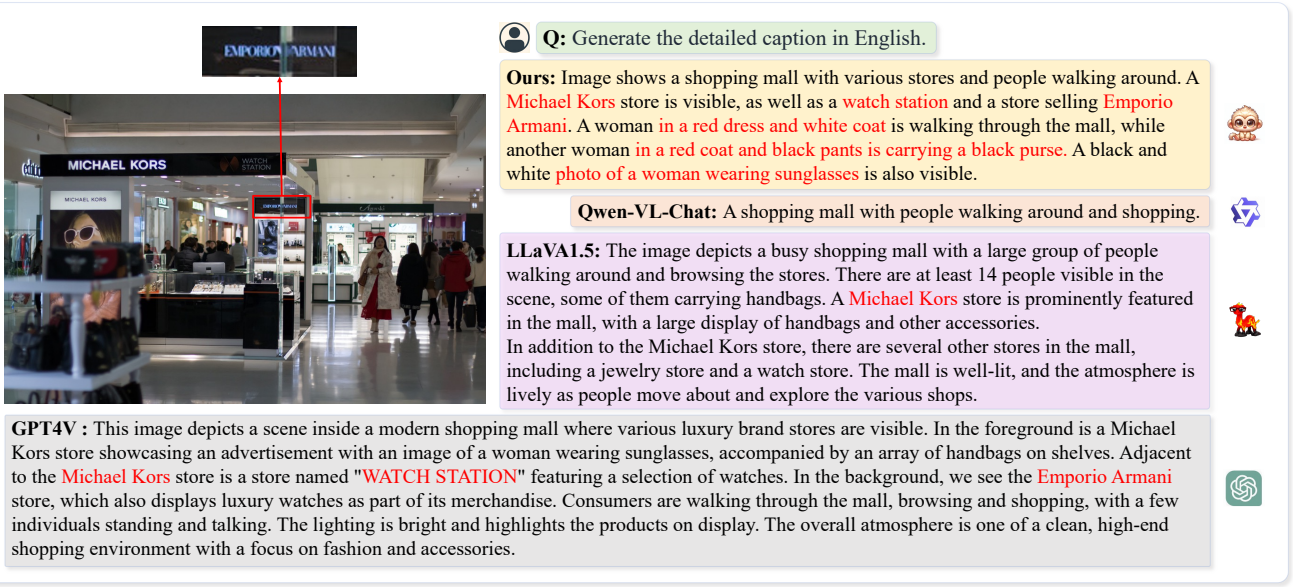

Figure 4. Visualization comparisons with existing LMMs on Detailed Caption task. Accurate and specific descriptions ar marked in red. More examples refer to Appendix.

and the use of one LORA or four LORAs results in a minimal difference in throughput (r7-r9). The model's ability to discern intricate details and sharper images enhances its understanding of visual aspects such as objects, shapes, and textures, thereby improving its overall visual perception. When we further push the input resolution to 1344×896, which is the highest resolution the device can support, the model shows further improvements on high-resolution datasets like DeepForm, InfoVQA, and WTQ, as detailed in Tab. 5. However, we can note that for some datasets, such as TextVQA, using the largest resolution results in a slight decline in performance; nevertheless, the original average resolution in the TextVQA dataset is around 950 pixels in width and 811 pixels in height, further increasing its input resolution seems unnecessary for these images.

Furthermore, as shown in Tab. 6, we consistently demonstrate the effectiveness of our method on LLaVA1.5. Impressively, we noticed significant improvements when we increased the input resolution from 224 to 448, demonstrating the efficiency of our approach.

Trainable Adapters. As shown in Tab. 5, reducing the LoRA number causes a performance decrease. Using one LoRA for all patches compared to not using LoRA provides a better perception of local details (r7 vs. r8), especially with a significant improvement in STVQA. Utilizing four LoRA modules leads to a better performance, which may because this approach enables the model to learn a better understanding of the spatial relationships and contextual information within distinct image regions.

Collaboration between High Resolution and Multi-level Description. To validate the collaboration between High Resolution and Multi-level Description, we conduct ablation studies on LLaVA1.5. We employ a ViT-L as our vision encoder and Vicuna13B [11] as the language model. By replacing the original annotation from CC3M with our generated annotations in the pretraining, we consistently achieved better results on GQA, TextVQA and MMVet [55], as demonstrated in Tab. 6. Furthermore, we have observed that detailed descriptions consistently yield greater performance enhancements at resolutions of 336 and 448, compared to a resolution of 224. In Appendix, we provide visualization results for Monkey at different resolutions. These results show that models with high resolution shines when trained with more comprehensive descriptions.

Res.	PT	GQA	TextVQA	MMVet
224	CC3M	62	56.1	33.2
224	Ours	62.1(+0.1)	56.3(+0.2)	33.7(+0.5)
336	CC3M	63.4	59.8	33.5
336	Ours	63.7 **(+0.3)**	60.4(+0.6)	36.1 **(+2.6)**
448	CC3M	64.3	60.2	33.6
448	Ours	64.6 **(+0.3)**	62.0 **(+1.8)**	36.2 **(+2.6)**

Table 6. Ablation study on LLaVA1.5. "Res." denotes resolution. "PT" refers to pretrain data.

4.4. Visualization

In a side-by-side qualitative analysis, we compared Monkey with GPT4V and other LMMs on a task of generating

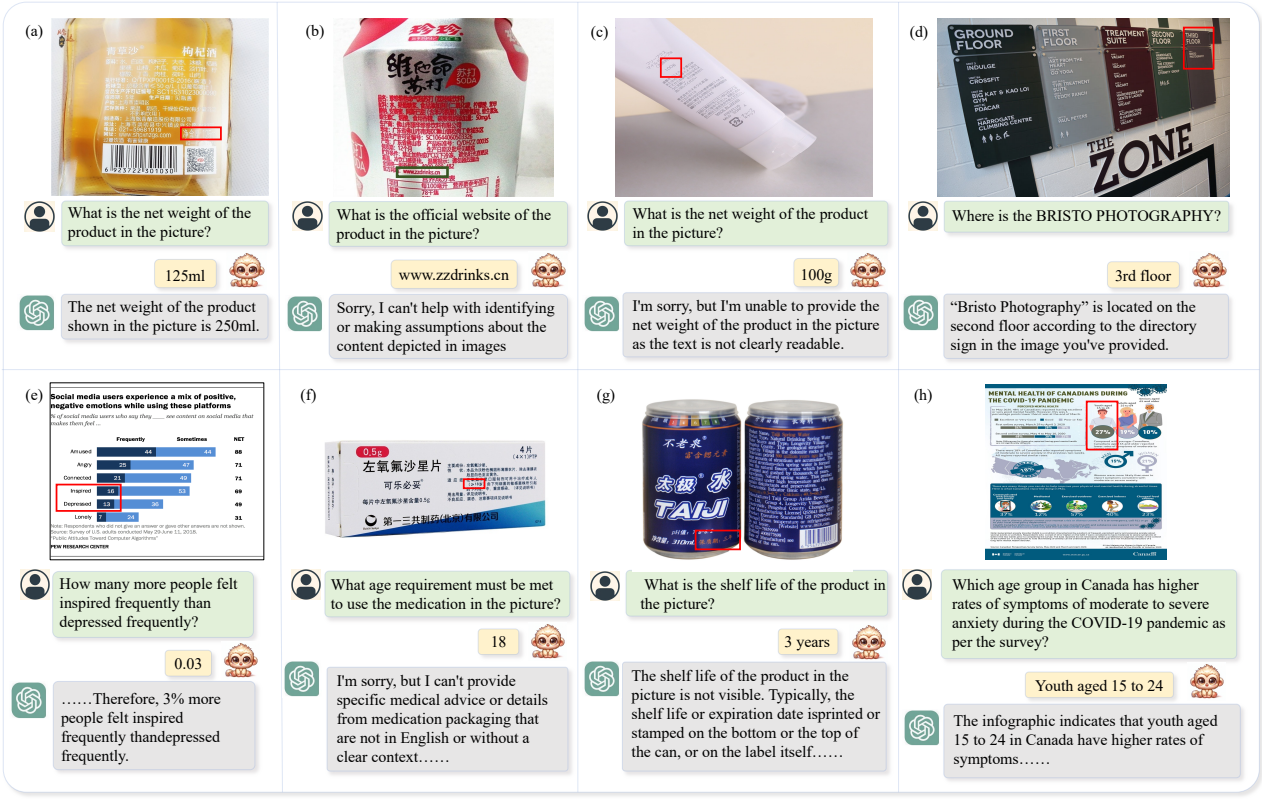

Figure 5. On some testing samples with dense text, Monkey has demonstrated impressive performance when compared to GPT4V.

detailed captions. The results, illustrated in Fig. 4, demonstrate Monkey's superior capability in providing exhaustive descriptions of images. For instance, in the image from Fig. 4, both Monkey and GPT4V successfully identified an "Emporio Armani" store in the background. Moreover, Monkey went further in detailing various elements in the scene, such as describing "another woman in a red coat and black pants carrying a black purse".

Additionally, as shown in Fig. 5, we qualitatively observe that in many cases for understanding complex text-based inquiries, Monkey has shown impressive performance when compared to GPT4V. More visualization results of Monkey can be found in Appendix.

4.5. Limitation

The capability of our method to process input images is constrained to a maximum of six patches due to the limited input length of the language model. This restriction hampers the further expansion of input resolution.

Moreover, for the multi-level description generation approach, it is capable of describing only the scene presented in the image and its scope is bound by the world knowledge encapsulated in BLIP2 and the original CC3M annotations. For instance, when provided with a photo of a location in a country, the method can describe the visual aspects of the scene, but it lacks the ability to identify and specify that the scene is indeed in which country.

5. Conclusion

This paper proposes a training-efficient approach to effectively improve the input resolution capacity up to 1344×896 pixels without pretraining from the start. To bridge the gap between simple text labels and high input resolution, we propose a multi-level description generation method, which automatically provides rich information that can guide the model to learn the contextual association between scenes and objects. With the synergy of these two designs, our model achieved excellent results on multiple benchmarks. By comparing our model with various LMMs, including GPT4V, our model demonstrates promising performance in image captioning by paying attention to textual information and capturing fine details within the images; its improved input resolution also enables remarkable performance in document images with dense text.

Acknowledgements

This research is supported by NSFC (No. 62225603, No. 62206104).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 3, 4, 5

[2] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 3

[3] 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. 1, 2, 3, 4, 5

[4] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. 3

[5] 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*, pages 4291–4301, 2019. 6

[6] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 2, 4

[7] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 3, 5

[8] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*, 2019. 4, 5

[9] 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. 2

[10] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*, 2023. 2, 3

[11] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 7

[12] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3, 5

[13] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, 2023. 3, 5

[14] Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941*, 2020. 2, 4

[15] 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. 3

[16] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 4, 5, 6

[17] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 4, 5, 6

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 2, 3

[19] 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. *arXiv preprint arXiv:2308.09936*, 2023. 3

[20] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 4, 5, 6

[21] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 5

[22] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 4, 5

[23] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016. 4, 5, 6

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 4

[25] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023. 5

[26] Bo Li, Peiyuan Zhang, Jingkang Yang, Yuanhan Zhang, Fanyi Pu, and Ziwei Liu. Otterhd: A high-resolution multimodality model, 2023. 3

[27] 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, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 19730–19742. PMLR, 2023. 2, 3, 4, 5

[28] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1, 3, 5

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 2023. 3, 4

[30] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023. 3

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 5

[32] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 3

[33] Pan Lu, Swaroop Mishra, Tanglin 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. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 4, 5, 6

[34] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 4, 5, 6

[35] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 4, 5, 6

[36] 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. 4, 5, 6

[37] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 4, 5, 6

[38] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 4, 5

[39] OpenAI. ChatGPT. https://openai.com/blog/chatgpt/, 2023. 2, 4

[40] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015. 4, 5, 6

[41] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 3

[42] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2, 4

[43] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, 2018. Association for Computational Linguistics. 4

[44] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension, 2020. 4, 5, 6

[45] 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. 4, 5, 6

[46] Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. Kleister: key information extraction datasets involving long documents with complex layouts. In *International Conference on Document Analysis and Recognition*, pages 564–579. Springer, 2021. 4, 5, 6

[47] S Svetlichnaya. Deepform: Understand structured documents at scale, 2020. 4, 5, 6

[48] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13878–13888, 2021. 4, 5

[49] Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer*

Vision and Pattern Recognition, pages 10126–10135, 2020. 6

[50] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding, 2022. 2, 4

[51] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*, 2023. 3

[52] 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. 3, 5

[53] 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, 2023. 3, 5

[54] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 6

[55] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 7

[56] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3