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

1. Summary of the Evaluation Benchmarks.

We present a comprehensive overview of the evaluation benchmarks utilized, along with their corresponding metrics in Tab. 1. For the Image Caption task, we selected two datasets: Flickr30K [24], which is an image caption dataset for natural images, and TextCaps [18], which is an image caption dataset for natural images with text. For general Visual Question Answering (VQA), we chose five commonly used datasets. VQAV2 [4] is an open-ended VQA dataset focused on natural images, while OKVQA [11] requires additional world knowledge. GQA [6] is a dataset designed for real-world visual reasoning and compositional question answering. ScienceQA [10] involves multimodal multiple-choice VQA on science topics, and VizWiz [5] aims to answer questions from blind individuals. In the domain of Scene Text-centric VQA, our selection includes TextVQA [19], AI2Diagram [7], STVQA [2], and ESTVQA [22]. AI2D is a multiple-choice VQA dataset that focuses on science diagrams, while the others involve reading and reasoning about text in natural images. For the STVQA and ESTVQA datasets, we followed the split provided by [9]. Regarding Doc-oriented VQA, we encompass various document images, including documents, charts, infographics, reports, and HTML tables. In the case of DeepForm [21] and KLC [20], we transform the Key Information Extraction task into a Visual Question Answering (VQA) task. Additionally, we evaluate Monkey on the MME benchmark [3], which measures perception and cognition abilities. Furthermore, for the ablation study on LLaVA1.5 [8], we adhere to the evaluation settings specified by LLaVA1.5.

Task	Dataset	Description	Split	Metric
Image Caption	Flickr30K [24]	Image caption for natural images	karpathy-test	CIDEr(↑)
	TextCaps [18]	Image caption for natural images with text	val	CIDEr(↑)
General VQA	VQAv2 [4]	Open-ended VQA about natural images	val	VQA Score(↑)
	OKVQA [11]	VQA involving world knowledge on natural images	val	VQA Score(↑)
	GQA [6]	Real-world visual reasoning and compositional question answering	test-dev	Accuracy(↑)
	ScienceQA [10]	Multimodal multiple choice VQA on science topics	test	Accuracy(↑)
	VizWiz [5]	Answering visual questions from blind people	val	VQA Score(↑)
Scene Text-centric VQA	TextVQA [19]	VQA involving reading and reasoning about text	val	VQA Score(↑)
	AI2Diagram [7]	Multiple choice VQA on science diagrams	test	Accuracy(↑)
	STVQA [2]	VQA involving reading and reasoning about text	test*	ANLS(↑)
	ESTVQA [22]	VQA involving reading and reasoning about text	test(English)*	ANLS(↑)
Doc-oriented VQA	DocVQA [13]	VQA on document images	test	ANLS(↑)
	ChartQA [12]	VQA on charts with visual and logical reasoning	test	Relaxed Accuracy(↑)
	InfoVQA [14]	VQA on infographic images	test	ANLS(↑)
	DeepForm [21]	Key Information Extraction on charity organizations' reports	test	Accuracy(↑)
	KLC [20]	Key Information Extraction on documents related to election spending	test	Accuracy(↑)
	WTQ [16]	VQA on semi-structured HTML tables sourced from Wikipedia	test	Accuracy(↑)
Evaluation Benchmark	MME [3]	Evaluation benchmark measuring perception and cognition abilities	Perception	Accuracy (↑)

Table 1. Summary of the evaluation benchmarks.

2. More Visualization Results

Figure 1. Visualization results.

We presented additional visualization results, where Fig. 1 demonstrates Monkey's capabilities in various VQA tasks. Monkey analyzes the question, identifies the key elements in the image relevant to answering the question, and exhibits the ability to perceive even minute text within the image. Moreover, Monkey can reason about the objects present in the scene and possesses a strong understanding of visual charts. In addition, Fig. 1 also showcases Monkey's impressive captioning ability, accurately describing various objects in the image and providing appropriate summaries.

3. More Examples of our Generated Data

Figure 2. Detailed captions generated by us.

In Fig. 2, we present the detailed captions generated by our method. Compared to the original annotations from the CC3M [17], our generated descriptions cover many more details of the image, providing a more detailed description of the image.

4. Comparison with other LMMs.

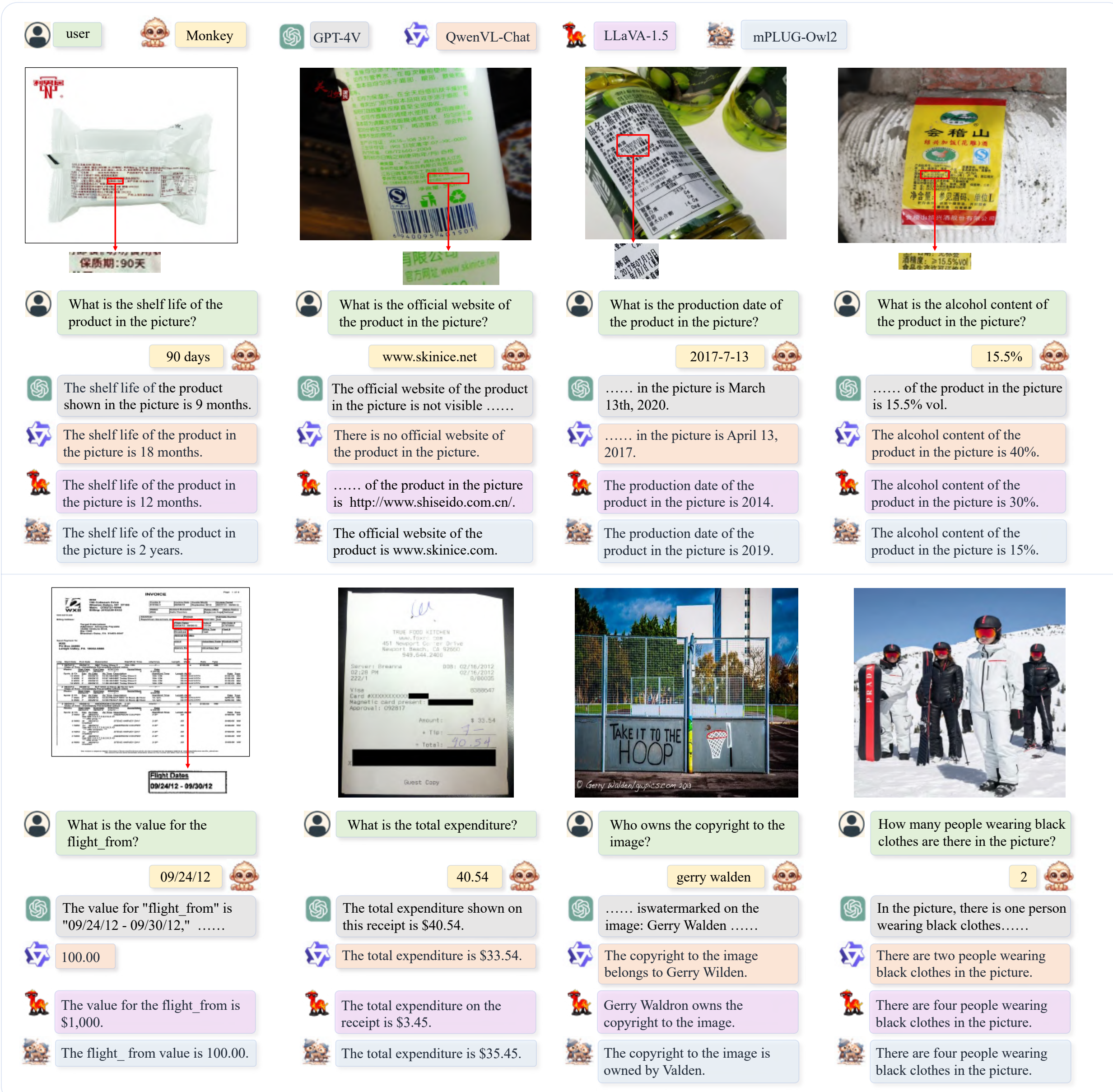

Figure 3. Visualization of Monkey's comparison with GPT-4V, QwenVL-Chat, LLaVA-1.5, and mPLUG-Owl2 on VQA task.

The comparison results of the VQA task in Fig. 3 indicate that after applying our method of scaling up the model size, Monkey has achieved significant performance advantages in tasks related to dense text. It not only surpasses the performance of QwenVL-Chat [1], LLaVA-1.5 [8], and mPLUG-Owl2 [23] but also achieves promising results compared to GPT-4V [15] in tasks related to dense text. This clearly demonstrates the importance of scaling up the model size for performance improvement in multimodal large models. It further validates the effectiveness of our method in enhancing the performance of multimodal large models.

In Fig. 4, the comparison between Monkey and GPT-4V, QwenVL-Chat, LLaVA-1.5, and mPLUG-Owl2 on Detailed Caption task is shown. It can be observed that Monkey accurately describes the content of the image and exhibits high sensitivity to the text within the image. It provides detailed descriptions of the image while ensuring accuracy.

user | Monkey | GPT-4V | QwenVL-Chat | LLaVA-1.5 | mPLUG-Owl2

user: Generate the detailed caption in English.

Monkey: This is an image of a painting of the Eiffel Tower in Paris. The painting is colorful and vibrant, with a blue and yellow sky in the background. There are also several other objects in the painting, including a yellow and blue building, a red and yellow flower, and a blue and yellow star. The painting is signed by the artist Aja.

GPT-4V: This is a vibrant and colorful impressionist painting of the Eiffel Tower in Paris. The artwork is characterized by bold, dynamic brush strokes and a vivaciously expressive color palette. The sky is a swirl of deep blues and bright yellows, reminiscent of Van Gogh's Starry Night, with hints of green and light blue creating a sense of motion and energy. Below the swirling sky, there is a dramatic sunset with vivid shades of pink, orange, and yellow blending into each other, evoking the feeling of a fiery Parisian evening. The Eiffel Tower itself is painted in warm yellow and orange tones, contrasting with the cool colors of the sky and providing a strong central focal point for the composition. In the foreground, there's a bridge over the Seine River with silhouettes of lampposts and trees lining the walkway, while the reflection on the water suggests calmness amidst the dynamism of the textures and colors above. The painting is signed "Aja '15" in the lower left corner, indicating the artist and year of creation.

QwenVL-Chat: Eiffel Tower Starry Night, Paris, France, original oil painting on canvas, palette knife textured modern Impressionism, by Ana Maria Edulescu

LLaVA-1.5: The image is a painting of the Eiffel Tower at night, with a colorful sky in the background. The tower is prominently featured in the center of the painting, with its height and grandeur emphasized. The sky is filled with vibrant colors, creating a captivating and lively atmosphere. There are several people scattered throughout the painting, likely enjoying the view of the tower and the beautiful sky. Some of them are closer to the tower, while others are further away, adding depth and interest to the scene. The combination of the tower, the sky, and the people creates a captivating and dynamic painting.

mPLUG-Owl2: The image is a painting of the Eifel tower, a famous landmark in Paris, France. The tower is prominently featured in the center of the painting, with its distinctive shape and structure. The sky in the background is painted in a vibrant and colorful manner, creating a lively and dynamic atmosphere. The painting captures the essence of the Eiffel tower and the beauty of the Parisian skyline.

user: Generate the detailed caption in English.

Monkey: A sign at an airport with various directions written in different languages, including English, Chinese, and Thai. The sign has arrows pointing to different areas such as baggage claim, immigration, and transfer counters. There are also various signs with different colors and shapes, including a yellow sign arrow pointing to the right.

GPT-4V: The image shows an airport information signboard with directions and labels in multiple languages including English, Chinese, and Thai. The top part of the sign displays arrows pointing to the left for "Immigration" and to the right for "Baggage Claim." Below this, the sign is divided into two sections, with the left side indicating "Transfer Counter Area" and the right side divided into "EAST" and "WEST" directions. Each section includes floor indicators, with the transfer counter and gates located on the 3rd floor. The gate areas are further divided with colorful squares labeled A, B, C, D (gates 1-10) and D, E, F, G (gates 1-6) with corresponding arrows indicating their direction. The bottom part of the sign provides information on transfers to various Thai cities such as Chiangmai, Chiangrai, Phuket, Krabi, Samui, and HatYai, accompanied by Chinese translations. The layout is clear, providing straightforward navigation for travelers within the airport.

QwenVL-Chat: An airport sign showing the way to different gates at the airport.

LLaVA-1.5: The image showcases a large signboard with directions and information for travelers. The signboard is written in both English and Chinese, catering to a diverse audience. It is divided into sections, each displaying various destinations and their corresponding gate numbers. Numerous arrows are pointing towards the different destinations, making it simple for travelers to navigate the airport. The signboard is situated in a public area, most likely within an airport, to help passengers find their way to their desired locations.

mPLUG-Owl2: The image features a sign with various directions and arrows pointing in different directions. The sign is likely to be found in an airport or a transportation hub, providing guidance to travelers. The arrows are arranged in a way that makes it easy for people to understand the different routes and destinations.

Figure 4. Visualization of Monkey's comparison with GPT-4V, QwenVL-Chat, LLaVA-1.5, and mPLUG-Owl2 on Detailed Caption task. Accurately described text is marked in red, while inaccurately described text is marked in blue.

5. Visualization results for models at different resolutions.

Figure 5. Visualization results of the VQA task at resolutions of 896, 784, and 672 respectively.

In Fig. 5, we performed VQA tasks testing at three different resolutions: 896, 784, and 672. The visual results obtained further validate the importance of our size expansion method for improving the performance of LMMs. While using a resolution of 896 for VQA tasks testing yielded correct results, using resolutions of 784 and 672 resulted in errors, with the smallest size of 672 showing more errors.

In Fig. 6, we conducted tests at three different resolutions: 896, 784, and 672. It can be observed that as the resolution decreases, the details in the images gradually become less visible to the model.

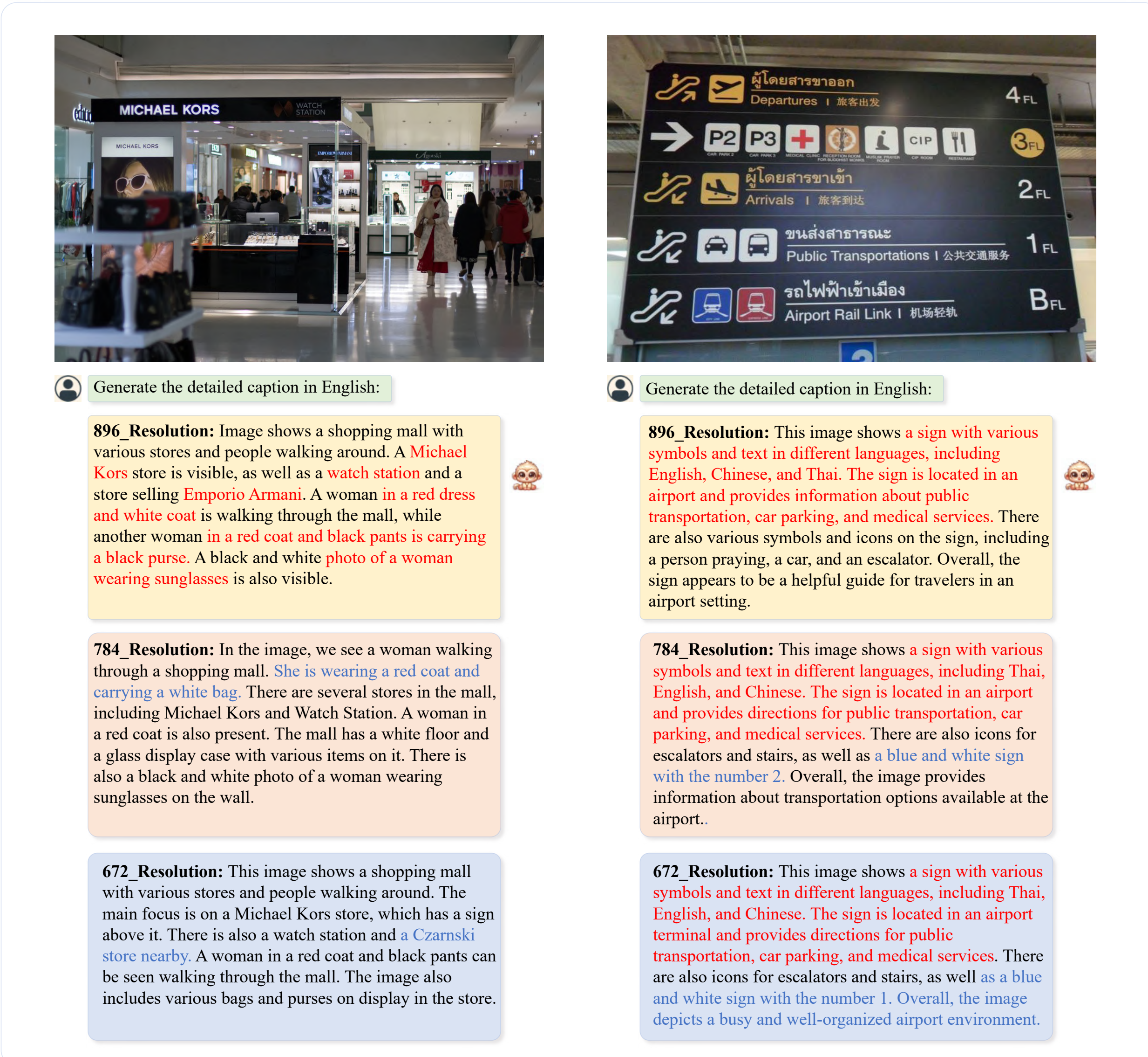

Figure 6. Visualization results of the detailed caption task at resolutions of 896, 784, and 672 respectively. Accurately described text is marked in red, while inaccurately described text is marked in blue.

6. Data Generation.

Hyperparameter Control in Data Generation Pipeline. The appropriate selection of hyperparameters is crucial. We empirically selected them based on qualitative results, finding SAM's default threshold and a 0.5 Image-Text Matching Score to be effective. We conducted a quantitative validation on 80 samples using the GPT-4V evaluation. The results shown in Tab. 2 reveal that SAM's threshold is relatively robust, and the 0.5 threshold for Image-Text Matching Score offers a better performance.

Pred-IOU-Thresh of SAM	0.4	0.6	0.88 (default)
GPT-4V Score	6.388	6.425	6.625
Image-Text Matching Score	0.2	0.5	0.7
GPT-4V Score	5.825	6.625	6.550

Table 2. Hyperparameter Control.

Comparison with LLaVA's GPT4 Method. While the GPT4 method in LLaVA is dependent on using manually annotated captions from the COCO dataset as a foundational basis for data generation, our approach focuses on generating original, detailed captions autonomously. Additionally, our detectors are skilled in revealing a spectrum of details in images, from text to nuanced object characteristics, which enables to enrich unlabeled data by extracting complex, multi-level details, paving the way for the creation of both cost-effective and accurate image descriptions.

Why choose BLIP2? We found that the performance is very similar in the GPT-4V evaluation when utilizing brief descriptions of local areas from other VLMs, as shown in Tab. 3. However, for generating approximately 5M descriptions, using BLIP2 takes approximately 3 days, while LLaVA and mPLUG-Owl require about 21 days and 32 days, respectively. For the sake of saving time, we choose BLIP2.

Model	LLaVA	mPLUG-Owl	Blip2
GPT-4V Score	6.663	6.225	6.625

Table 3. Performance of Different LMM.

7. Ablation study on Global Feature.

We conducted experiments on the presence or absence of global features at a resolution of 896. By adding global features, the results showed a 7.5% performance gain on TextVQA, a 0.6% performance gain on GQA, and a 6.2% performance gain on DocVQA. This demonstrated that global features contribute to enhancing the overall performance.

References

[1] 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. 4

[2] 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. 1

[3] 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. 1

[4] 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. 1

[5] 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. 1

[6] 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. 1

[7] 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. 1

[8] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1, 4

[9] 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. 1

[10] 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. 1

[11] 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. 1

[12] 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. 1

[13] 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. 1

[14] 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. 1

[15] OpenAI. Gpt-4 technical report, 2023. 4

[16] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015. 1

[17] 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. 3

[18] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension, 2020. 1

[19] 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. 1

[20] 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. 1

[21] S Svetlichnaya. Deepform: Understand structured documents at scale, 2020. 1

[22] 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. 1

[23] 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. 4

[24] 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. 1