DocThinker: Explainable Multimodal Large Language Models with Rule-based Reinforcement Learning for Document Understanding

Supplementary Material

1. Prompt Template

As illustrated in Tab. 1, the prompt template is designed to instruct the MLLM to produce structured output o, which includes both a reasoning trace and a final output encoded in designated XML-like tags (<think>...</think>and <answer>...</answer>).

2. Rewiew of GRPO

The Group Relative Policy Optimization (GRPO) algorithm, first introduced in DeepSeekMath [17], is a reinforcement learning framework designed to improve reasoning without the need for a separate critic model, a key limitation of existing methods such as Proximal Policy Optimization (PPO)[15]. Traditional RL approaches like PPO rely on a value network to estimate the quality of model predictions, which can introduce instability and additional computational costs. In contrast, GRPO directly compares a group of generated responses, making it a more efficient alternative for large-scale language model training.

In GRPO, given a question q, the old policy model $\pi_{\theta_{\text{old}}}$ first generates a group of different candidate response outputs $\{o_1, o_2, ..., o_G\}$ with size of G. These response outputs are then evaluated through a rule-based reward function $R(q, o)$ to obtain G rewards denoted as $\{r_1, r_2, ..., r_G\}$ correspondingly, which is defined as follows:

$$r_i = R(q, o_i) = \begin{cases} 1, & \text{if } o_i = \text{ground truth}, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

where $R(\cdot, \cdot)$ is the rule-based verifiable reward function. R takes the question and output pair (q, o_i) as inputs, and checks whether the prediction o_i is correct compared to ground truth under predefined rules. In our works, we proposed multi-objective reward functions tailored for document understanding, to incentivize the model to generate human-understandable reasoning steps, while ensuring robust generalization across diverse document types and tasks.

Instead of computing absolute values for each response, GRPO normalizes the rewards within the group, ensuring that the model learns from relative advantages. Specifically, the advantage is computed by taking the difference between each reward and the $mean$ of the group, normalized by the standard deviation std, formulated as follows:

$$A_i = \frac{r_i - \text{mean}(\{r_1, \ldots, r_G\})}{\text{std}(\{r_1, \ldots, r_G\})}, \tag{2}$$

where A_i represents the advantage of i-th output o_i, meaning the relative quality of the i-th responses. The advantage A_i is sequence-level normalized reward, and we set the advantage $A_{i,t}$ of t-th auto-regressive decoding time step token in the output o_i as the sequence-level advantage A_i. This process eliminates the need for a critic network, making policy updates computationally efficient and stable. The intuition behind GRPO objective is to maximize the advantage of the generated responses, while ensuring that the model remains close to the reference policy model π_{ref}. Consequently, the GRPO loss $\mathcal{L}_{\text{GRPO}}$ is defined as follows:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left[\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\varphi\left[\pi_\theta(o_{i,t} \mid q, o_{i,<t})\right]}A_{i,t} - \beta\mathbb{D}_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}})\right], \tag{3}$$

where the first term represents the scaled advantage and the second term is regularization to penalize deviations from the reference policy π_{ref} through Kullback–Leibler (KL) divergence $\mathbb{D}_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}})$, helping prevent catastrophic forgetting. $\varphi[\cdot]$ represents stop gradient operation. θ is the trainable parameter of the current policy model π_θ. $\beta \in \mathbb{R} \geq 0$ is a hyper-parameter and controls the regularization strengths. GRPO encourages the model to favor better answers with a high reward value within the group.

In the original DeepSeekMath [17] paper, the objective $\mathcal{L}_{\text{GRPO}}$ formulation in Eq. (3) is generalized to account for multiple updates after each group response generation by leveraging the clipped surrogate objective to ensure that updates do not deviate excessively from the reference policy by bounding the policy ratio between $1-\epsilon$ and $1+\epsilon$ via $\text{clip}(\cdot, 1-\epsilon, 1+\epsilon)$ function, formulated as follows:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left[\min\left(\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}A_{i,t}, \text{clip}\left(\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}, 1-\epsilon, 1+\epsilon\right)A_{i,t}\right) - \beta\mathbb{D}_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}})\right], \tag{4}$$

where $\epsilon \in \mathbb{R} \geq 0$ is a clipping-related hyper-parameter introduced in PPO [15] for stabilizing training by preventing drastic changes in policy updates. In practice, as in the original paper, we only do one update per generation. In this condition, π_θ is equal to $\pi_{\theta_{odd}}$, so we can simplify the loss to the first form defined in Eq. (3).

The prompt template

You are given an original question. Your task is to provide an accurate answer to the question and determine the bounding box coordinates of the region that best supports your answer.

To enhance clarity and interpretability, you should:
- Understand the intent behind the original question.
- Modify the original question by adding relevant descriptive phrases and details based on the provided image.
- Ensure that the modified question remains semantically similar to the original.

Your response has two parts:
1. **Thinking Process:** Before outputting the answer, describe your reasoning process within <think></think>tags.
2. **Final Output:** Provide the answer in JSON format within <answer></answer>tags. The JSON should contain the following keys:
    - **rephrase_question**: The improved and more descriptive version of the original question.
    - **bbox_2d**: The bounding box coordinates [x_min, y_min, x_max, y_max] of the region that supports the answer.
    - **final_answer**: The actual answer to the question.

Example Output Format:
Original question: "What is the man doing?"
<think>
reasoning process here
</think>
<answer>
{
    "rephrase_question": "What is the man wearing while preparing to shoot the basketball near the hoop?",
    "bbox_2d": [150, 300, 400, 600],
    "final_answer": "answer here."
}
</answer>

Original question: "{Question}"

Table 1. **The template of our employed prompt for DocThinker.** Question will be replaced with the specific question during training and inference.

In practice, KL divergence is estimated using the unbiased estimator introduced by [14]. The approximator is defined as follows:

$$\mathbb{D}_{\mathrm{KL}}\left[\pi_\theta \| \pi_{\mathrm{ref}}\right] = \frac{\pi_{\mathrm{ref}}\left(o_{i,t} \mid q, o_{i,<t}\right)}{\pi_\theta\left(o_{i,t} \mid q, o_{i,<t}\right)} - \log \frac{\pi_{\mathrm{ref}}\left(o_{i,t} \mid q, o_{i,<t}\right)}{\pi_\theta\left(o_{i,t} \mid q, o_{i,<t}\right)} - 1, \quad (5)$$

where this approximation ensures that KL estimates remain positive and computationally stable throughout training.

3. More Results and Analysis

RoI Detection Results. Tab. 2 presents the RoI detection performance, measured by Top-1 Accuracy@0.5, across multiple document understanding and general reasoning tasks. A higher score indicates better alignment between the model's predicted bounding boxes and the ground truth key regions annotated in the Visual CoT benchmark [16]. Compared to VisCoT-7B, our model DocThinker-7B (336^2) achieves substantial improvements across all tasks, particularly in document-oriented datasets. It outperforms the strongest baseline by a large margin on DocVQA (38.3 vs. 20.4), TextCaps (58.6 vs. 46.3), and TextVQA (59.2 vs. 57.6). More challenging datasets, such as DUDE and

Document-oriented Understanding								
			Doc/Text					Chart
MLLM	Res.	Strategy	DocVQA	TextCaps	TextVQA	DUDE	SROIE	InfoVQA
VisCoT-7B [16]	224^2	SFT	13.6	41.3	46.8	5.0	15.7	7.2
VisCoT-7B [16]	336^2	SFT	20.4	46.3	57.6	9.6	18.5	10.0
DocThinker-7B	336^2	RL	**38.3**	**58.6**	**59.2**	**27.5**	**32.1**	**23.6**

General Multimodal Understanding								
			General VQA		Relation Reasoning			Average
MLLM	Res.	Strategy	Flickr30k	Visual7W	GQA	Open Images	VSR	
VisCoT-7B [16]	224^2	SFT	49.6	31.1	42.0	57.6	69.6	37.2
VisCoT-7B [16]	336^2	SFT	51.3	29.4	49.5	59.3	54.0	37.6
DocThinker-7B	336^2	RL	**55.7**	**36.3**	**53.6**	**67.1**	**59.8**	**46.5**

Table 2. Detection performance (Top-1 Accuracy@0.5) on the Visual CoT benchmark [16]. Grey results indicate zero-shot performance. Res. shorts for image resolution. Average refers to the average accuracy across eleven datasets. The ground truth bounding boxes used for computing the metric are the intermediate CoT bounding boxes annotated in the Visual CoT benchmark.

Method	Scene Text-Centric VQA		Document-oriented VQA			KIE		
	STVQA	TextVQA	DocVQA	InfoVQA	ChartQA	FUNSD	SROIE	POIE
BLIP2-OPT-6.7B [8]	20.9	23.5	3.2	11.3	3.4	0.2	0.1	0.3
mPLUG-Owl [20]	30.5	34.0	7.4	20.0	7.9	0.5	1.7	2.5
InstructBLIP [3]	27.4	29.1	4.5	16.4	5.3	0.2	0.6	1.0
LLaVAR [22]	39.2	41.8	12.3	16.5	12.2	0.5	5.2	5.9
BLIVA [6]	32.1	33.3	5.8	23.6	8.7	0.2	0.7	2.1
mPLUG-Owl2-8 [21]	49.8	53.9	17.9	18.9	19.4	1.4	3.2	9.9
LLaVA1.5-7B [11]	38.1	38.7	8.5	14.7	9.3	0.2	1.7	2.5
TGDoc [18]	36.3	46.2	9.0	12.8	12.7	1.4	3.0	22.2
UniDoc [4]	35.2	46.2	7.7	14.7	10.9	1.0	2.9	5.1
DocPedia [5]	45.5	60.2	47.1	15.2	46.9	29.9	21.4	39.9
Monkey-8B [9]	54.7	64.3	50.1	25.8	54.0	24.1	41.9	19.9
InternVL-8B [2]	62.2	59.8	28.7	23.6	45.6	6.5	26.4	25.9
InternLM-XComposer2-7B [19]	59.6	62.2	39.7	28.6	51.6	15.3	34.2	49.3
TextMonkey-9B [13]	61.8	65.9	64.3	28.2	58.2	32.3	47.0	27.9
InternVL2-2B [1]	65.6	66.2	76.7	46.8	67.6	42.0	68.0	66.8
Mini-Monkey-2B [7]	67.2	68.8	78.4	50.0	67.3	43.2	70.5	71.2
DocThinker-7B	**68.4**	**69.7**	**78.8**	**52.3**	**67.8**	**47.2**	**73.1**	**72.8**

Table 3. Quantitative accuracy (%) comparison of DocThinker with existing multimodal large language models (MLLMs) on widely used benchmark. Following TextMonkey [13], we use the accuracy metrics to evaluate our method.

SROIE, which require precise text-region localization, also see significant gains, with our model scoring 27.5 and 32.1, compared to 9.6 and 18.5, respectively. Beyond document tasks, DocThinker demonstrates stronger generalization in VQA and relational reasoning benchmarks, outperforming VisCoT-7B in Flickr30k (55.7 vs. 51.3), GQA (53.6 vs. 49.5), and Open Images (67.1 vs. 59.3). The model also improves Visual7W and VSR performance, achieving 36.3 and 59.8, respectively. These results confirm that reinforcement learning with RoI-based rewards enhances the model's ability to precisely localize key regions, leading to better multimodal alignment and more reliable reasoning outputs. The superior results demonstrate DocThinker's effectiveness in both structured document reasoning and general multimodal comprehension.

OCRBench Results. To further evaluate DocThinker beyond the Visual CoT Benchmark [16], we assess its performance on OCRBench [12], a widely used benchmark for text-centric multimodal understanding. Following the TextMonkey [13] evaluation framework, we use accuracy metrics (%) across scene text-based VQA, document-oriented VQA, and key information extraction (KIE) tasks. As shown in Tab. 3, DocThinker-7B achieves state-of-the-art performance, surpassing previous MLLMs across all categories.

In scene text VQA, our model scores 68.4% on STVQA and 69.7% on TextVQA, outperforming Mini-Monkey-2B [7] and InternVL2-2B [1]. In document-oriented VQA, DocThinker reaches 78.8% on DocVQA, 52.3% on InfoVQA, and 67.8% on ChartQA, consistently leading across structured text understanding tasks. For key information extraction (KIE), which demands precise text localization and recognition, DocThinker sets new benchmarks with 47.2% on FUNSD, 73.1% on SROIE, and 72.8% on POIE, surpassing Mini-Monkey-2B and other strong baselines. These results highlight the effectiveness of reinforcement learning with structured rewards in improving both text-centric reasoning and document comprehension, demonstrating DocThinker's ability to generalize across complex multimodal text understanding tasks.

Accuracy of Rephrased Questions. We construct a new training set using model generated rephrased questions and fine-tuned on Qwen. As shown in Tab. 4, this model outperforms one trained on original QA pairs (0.548 vs. 0.497 average score on Visual CoT), demonstrating that rephrased questions preserve and even enhance task relevance.

Method	Res.	Data	Avg.
Qwen2.5VL-7B	336^2	Original QA	0.497
		Rehprase QA	**0.548**

Table 4. Accuracy of rephrased questions.

Hallucination in Rephrased Questions. Following HallusionBench [10], we use GPT-4 to judge 200 randomly sampled rephrased questions. As shown in Tab. 5, results show 96% correctness, 0% inconsistency, and 4% unclear, indicating that language hallucinations are rare. Besides, a human evaluation confirms 99% correctness.

	Semantic Consistency			Human Check
	Correct	Inconsistent	Unclear	
Rephrase question	96%	0%	4%	99%

Table 5. Hallucination in rephrased questions.

Hallucination of the Resulting Model. We evaluated hallucination rate of the resulting model on HallusionBench [10]. As shown in Tab. 6, our model achieves 69.8%, outperforming baseline Qwen2.5VL (69.4%).

Method	HallusionBench
Qwen2.5VL-7B	69.4%
Ours	**69.8%**

Table 6. Hallucination of the resulting model.

Scaling Effects. We scale training data from 4k to 64k using samples from Visual CoT. As shown in Fig. 1, the average results of DocThinker-7B (336^2) improve consistently with more data, demonstrating a clear scaling effect.

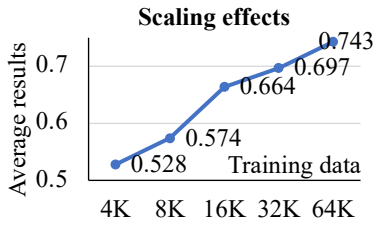

Figure 1. Scaling effects of DocThinker on document benchmarks. Increasing training data from 4k to 64k consistently improves performance, showing the scalability of GRPO.

Failure Case. As shown in Fig. 2, although the model correctly locates the 2D bounding box of RoI and predicts the final answer "In the Levant", the rephrased question incorrectly identifies the author as "Charles Dvdley Warner" instead of the correct "Charles Dudley Warner". This error likely stems from visual confusion caused by the stylized font on the book cover, where the letter "u" was misperceived as "v". While this hallucination does not affect the final answer, it reveals a limitation in the model's visual text recognition. Future work may improve robustness against such artistic distortions to reduce semantic hallucinations in intermediate outputs.

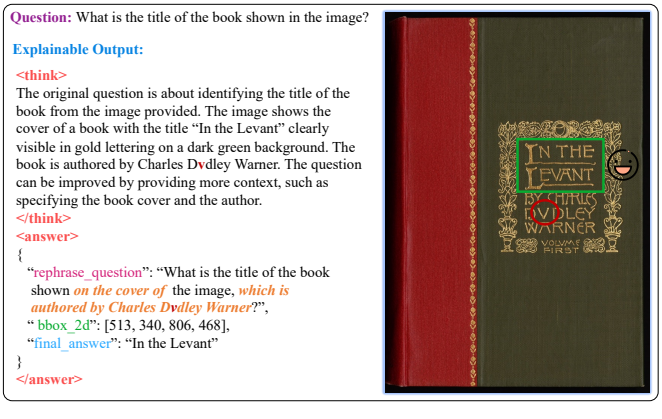

Figure 2. Failure case. Although the model correctly predicts the final answer "In the Levant" and localizes the 2D bounding box of RoI accurately, the rephrased question contains a hallucination: it misidentifies the author as "Charles Dvdley Warner" instead of "Charles Dudley Warner". This error likely results from visual confusion caused by the stylized font on the book cover, where the letter "u" was misread as "v". While the final output remains correct, this case highlights the model's vulnerability to artistic distortions in text recognition.

References

[1] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiao wen Dong, Hang Yan, Hewei Guo, Conghui He, Zhenjiang Jin, Chaochao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, and Yu Qiao. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 2024. 3, 4

[2] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Intern vl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 3

[3] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Neural Information Processing Systems*, 2023. 3

[4] Hao Feng, Zijian Wang, Jingqun Tang, Jinghui Lu, Wen gang Zhou, Houqiang Li, and Can Huang. Unidoc: A universal large multimodal model for simultaneous text detection, recognition, spotting and understanding. *ArXiv*, abs/2308.11592, 2023. 3

[5] Hao Feng, Qi Liu, Hao Liu, Jingqun Tang, Wengang Zhou, Houqiang Li, and Can Huang. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *Science China Information Sciences*, 67(12):1–14, 2024. 3

[6] Wenbo Hu, Y. Xu, Y. Li, W. Li, Z. Chen, and Z. Tu. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 3

[7] Mingxin Huang, Yuliang Liu, Dingkang Liang, Lianwen Jin, and Xiang Bai. Mini-monkey: Multi-scale adaptive cropping for multimodal large language models. In *ICLR*, pages 1–15, 2025. 3, 4

[8] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023. 3

[9] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26753–26763, 2024. 3

[10] Fuxiao Liu, Tianrui Guan, Xiyang Wu, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v(ision), llava-1.5, and other multi-modality models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 4

[11] 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. 3

[12] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 2024. 3

[13] 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*, abs/2403.04473, 2024. 3

[14] John Schulman. Approximating KL Divergence. `http://joschu.net/blog/kl-approx.html`, 2020. 2

[15] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017. 1

[16] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Neural Information Processing Systems*, 2024. 2, 3

[17] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300, 2024. 1

[18] Yonghui Wang, Wen gang Zhou, Hao Feng, Keyi Zhou, and Houqiang Li. Towards improving document understanding: An exploration on text-grounding via mllms. *ArXiv*, abs/2311.13194, 2023. 3

[19] Xiao wen Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *ArXiv*, abs/2401.16420, 2024. 3

[20] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yi Zhou, Junyan Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qiang Qi, Ji Zhang, and Feiyan Huang. mplug-owl: Modularization empowers large language models with multimodality. *ArXiv*, abs/2304.14178, 2023. 3

[21] Qinghao Ye, Haiyang Xu, Jiabo Ye, Mingshi Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owi2: Revolutionizing multi-modal large language model with modality collaboration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051, 2024. 3

[22] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 3