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

Wenwen Yu[1], Zhibo Yang[2], Yuliang Liu[1], Xiang Bai[1]✉
[1]Huazhong University of Science and Technology, [2]Alibaba Group
{wenwenyu,ylliu,xbai}@hust.edu.cn, yangzhibo450@gmail.com

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated remarkable capabilities in document understanding. However, their reasoning processes remain largely black-box, making it difficult to ensure reliability and trustworthiness, especially in high-stakes domains such as legal, financial, and medical document analysis. Existing methods use fixed Chain-of-Thought (CoT) reasoning with supervised fine-tuning (SFT) but suffer from catastrophic forgetting, poor adaptability, and limited generalization across domain tasks. In this paper, we propose DocThinker, a rule-based Reinforcement Learning (RL) framework for dynamic inference-time reasoning. Instead of relying on static CoT templates, DocThinker autonomously refines reasoning strategies via policy learning, generating explainable intermediate results, including structured reasoning processes, rephrased questions, regions of interest (RoI) supporting the answer, and the final answer. By integrating multi-objective rule-based rewards and KL-constrained optimization, our method mitigates catastrophic forgetting and enhances both adaptability and transparency. Extensive experiments on multiple benchmarks demonstrate that DocThinker significantly improves generalization while producing more explainable and human-understandable reasoning steps. Our findings highlight RL as a powerful alternative for enhancing explainability and adaptability in MLLM-based document understanding. Code will be available at https://github.com/wenwenyu/DocThinker.

1. Introduction

Multimodal Large Language Models (MLLMs) [29–31] have significantly advanced document understanding, yet their reasoning mechanisms remain largely opaque. This lack of explainability [10, 39] limits their application in high-stakes domains such as legal, financial, and medical document analysis, where transparency is critical for ensuring trustworthiness. Unlike human reasoning, which involves structured and multi-step inference, MLLMs typically operate as black-box systems, making it difficult to validate their decision-making process [4]. While Chain-of-Thought (CoT) [43] prompting has been widely adopted to enhance explainability, existing approaches heavily rely on static reasoning templates, which struggle to generalize across diverse complex scenarios and tasks.

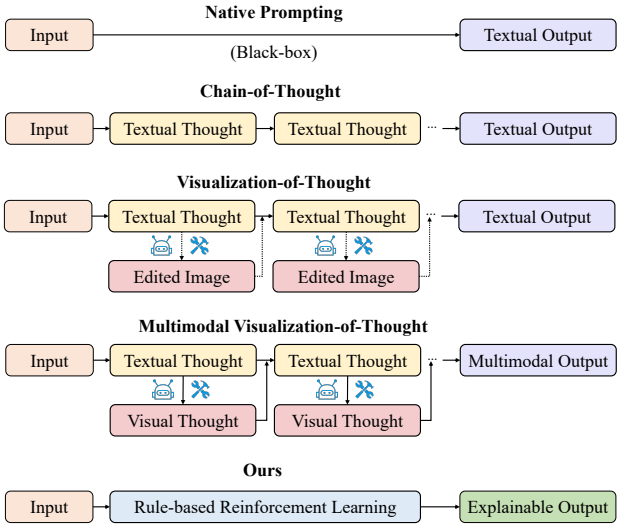

Figure 1. **Comparison of different approaches for improving model's explainability and transparency in MLLM-based document understanding.** Traditional methods, including Chain-of-Thought (CoT), Visualization-of-Thought (VoT), and Multimodal Visualization-of-Thought (MVoT) rely on static reasoning templates, predefined heuristics, or external agents and tools, limiting their adaptability and generalization. In contrast, the proposed DocThinker leverages rule-based reinforcement learning to explore diverse reasoning paths and generate explainable intermediate steps, including reasoning traces, rephrased questions, regions of interest (RoI) supporting the answer, and the final answer, enabling more adaptive and explainable document understanding.

To address this, recent research has explored multimodal CoT reasoning techniques, as shown in Fig. 1. ReFocus [8] presents a visual-editing-based CoT framework, allowing models to selectively highlight and modify key regions via

✉Corresponding author.

invoking external tools and agents in structured document images, improving comprehension of charts and tables. Visual CoT [35] introduces multi-turn processing pipelines that dynamically focus on key regions in visual images, enabling more explainable intermediate reasoning steps. Similarly, Multimodal Visualization-of-Thought (MVoT) [17] extends CoT reasoning by generating interleaved visual and textual reasoning traces, aiming to improve transparency. The Mind's Eye of LLMs [44] further proposes Visualization-of-Thought (VoT), a technique that elicits spatial reasoning by generating visual representations of thought processes. But this method applies in navigation-based applications, and its adaptability to document understanding remains limited. Besides, these approaches still depend on predefined heuristics and static reasoning paths, making them inherently inflexible and susceptible to catastrophic forgetting and poor generalization across varied document types and tasks.

Another emerging direction is reinforcement learning (RL)-based [40] reasoning, which has shown promise in overcoming the rigidity of fixed CoT methods. DeepSeek-R1 [5] framework demonstrates that pure RL training can incentivize emergent reasoning behaviors without relying on extensive supervised fine-tuning (SFT), achieving state-of-the-art performance in complex reasoning tasks. Inspired by this, MedVLM-R1 [32] applies RL techniques to medical vision-language models, proving its effectiveness in enhancing transparency and generalization in medical image understanding. Visual-RFT [24] introduces Visual Reinforcement Fine-Tuning, a reward-driven optimization framework designed to enhance the reasoning capabilities of vision-language models. Unlike conventional supervised fine-tuning, which relies on large annotated datasets, Visual-RFT employs verifiable reward functions to guide learning, significantly improving data efficiency. Experimental results indicate that reinforcement fine-tuning improves performance in open-vocabulary detection, few-shot object recognition, and reasoning grounding tasks, demonstrating its ability to generalize across diverse visual domains. While these methods primarily focus on general visual tasks, their success highlights the potential of RL in optimizing reasoning strategies for MLLMs. However, RL-based approaches for document understanding remain underexplored, particularly in designing effective reward functions that optimize both reasoning adaptability and explainability.

While humans naturally employ structured, multi-step reasoning when interpreting documents, integrating inference-time reasoning into MLLM-based document understanding is still an open challenge. Inspired by recent advancements, we propose DocThinker, a novel rule-based Reinforcement Learning (RL) framework designed for inference-time reasoning in document understanding. Unlike fixed CoT or VoT-style methods, DocThinker explores diverse reasoning paths and get explainable intermediate steps, including explicit reasoning traces, rephrased questions, regions of interest (RoI) supporting the answer, and the final answer, highlighting its ability to produce more flexible and varied outputs compared to traditional CoT. Instead of following fixed reasoning templates, DocThinker autonomously refines its reasoning process through policy learning based on the Group Relative Policy Optimization (GRPO) algorithm [36], mitigating catastrophic forgetting and improving adaptability. The model is trained using reinforcement learning with a proposed multi-objective reward function, enabling it to self-adapt to diverse document structures while preserving explainability. Although VoT-like methods generate grounded thought, they lack revision ability. Our RL enables self-reflection and correction that is complementary to VoT-like methods. Additionally, KL-constrained optimization is employed to ensure stable policy updates, preventing reward exploitation and preserving reasoning coherence.

Our main contributions are summarized as follows:

- We introduce DocThinker, to the best of our knowledge, the first RL-based framework for document understanding, enabling adaptive inference-time reasoning without relying on fixed CoT templates.
- We propose a set of multi-objective reward functions that incentivize the model to generate human-understandable reasoning steps, while ensuring robust generalization across diverse document types and tasks.
- We conduct extensive experiments on multiple benchmark datasets, demonstrating that DocThinker significantly improves generalization and explainability compared to existing CoT-based and SFT-based methods. Our findings highlight the potential of RL as a key enabler for more explainable, adaptable, and reliable MLLM-based document understanding systems.

2. Related Work

2.1. MLLMs for Document Understanding

Multimodal Large Language Models (MLLMs) have shown strong potential in document understanding by integrating textual and visual elements. Existing approaches enhance comprehension through layout-awareness, high-resolution processing, and specialized encoding techniques. DocLLM [41], LLaVAR [53], and mPLUG-DocOwl [46] improve text-centric document reasoning via instruction tuning, while methods like DocPedia [7] and Vary [42] refine image processing for structured text extraction. Other models, such as UReader [47] and InternVL1.5 [2], incorporate OCR and adaptive resolution techniques to enhance text recognition. To further optimize efficiency, models like TextMonkey [23] and Fox [19] introduce token compression and unified encoding for multi-page document analysis. Additional improvements, including compression strategies in DocKylin [52] and multi-scale integration in StrucTexTv3 [26], enhance

structure-aware reasoning. Despite these advancements, explainability remains a critical limitation. While step-wise reasoning frameworks [51] and visual editing-based approaches like ReFocus [8] improve transparency, they rely on static reasoning strategies that hinder adaptability. Most models depend on supervised fine-tuning (SFT), which improves task performance but often leads to overfitting and weak generalization across diverse document types. This highlights the need for adaptive learning frameworks capable of dynamic reasoning refinement while preserving explainability across varied document scenarios.

2.2. RL for Explainability and Reasoning

Explainability is crucial for deploying MLLMs in sensitive domains, ensuring transparency and trust in model decisions [4, 39]. Reinforcement Learning (RL) has emerged as a promising alternative to supervised fine-tuning (SFT) [3], addressing overfitting and limited generalization by allowing models to self-improve reasoning strategies through interaction with an environment. Unlike static Chain-of-Thought (CoT) approaches, RL enhances adaptability, explainability, and generalization in complex reasoning tasks. Recent advancements demonstrate RL's effectiveness in language and vision-language models. The OpenAI o1 model [13] applies RL to enhance reasoning capabilities, while DeepSeek-R1-Zero [5] achieves better reasoning and thinking process ability by training entirely with RL, incentivizing emergent reasoning via its Group Relative Policy Optimization (GRPO) framework without relying on SFT. MedVLM-R1 [32] extends this to medical image analysis, showing improved explainability and transparency. In multimodal learning, Visual-RFT [24] introduces verifiable reward functions, improving data efficiency and reasoning adaptability in open-vocabulary detection, few-shot recognition, and grounding tasks. RLHF-V [49] further aligns MLLMs with human trustworthiness through fine-grained RL-based feedback. Despite RL's success in vision-language tasks, its application in document understanding remains largely unexplored. Existing RL-based approaches fail to jointly process text, layout, and visual elements while maintaining explainability. Furthermore, designing effective reward functions that balance reasoning adaptability and explainability remains an open challenge. To bridge this gap, we propose DocThinker, a rule-based RL framework optimized with Group Relative Policy Optimization (GRPO) [36], incorporating verifiable multi-objective rewards to enable inference-time reasoning for complex document understanding.

3. Methodology

3.1. Preliminary

Group Relative Policy Optimization (GRPO). The GRPO algorithm, first introduced in DeepSeekMath [36], 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)[34]. 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, which will be detailed in Sec. 3.2.2, 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:

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

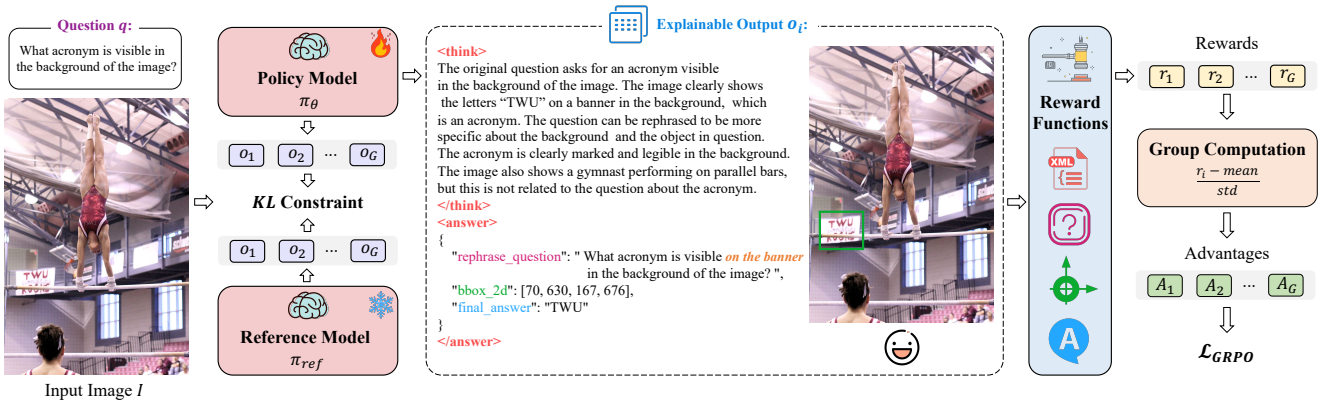

Figure 2. **Schematic illustration of the proposed DocThinker framework.** Given an input image I and question q, we first sample G candidate outputs $\{o_i\}_{i=1}^{G}$ from the old policy model $\pi_{\theta_{old}}$. The i-th output o_i has explainable and human-understandable intermediate results, including reasoning processes, rephrased question, regions of interest (RoI) supporting the answer, and final answer. Then we compute a reward r_i for each output o_i using our proposed multi-objective reward functions, which will be detailed in Sec. 3.2.2, including XML format reward, rephrased question reward, RoI IoU reward, and final answer accuracy reward. Subsequently, each reward r_i is normalized by subtracting the group average $mean$ and dividing by the group standard deviation std to get a group relative advantage A_i. Finally, we optimize the current policy model π_θ where θ is trainable parameters by maximizing the advantage A_i while ensuring that the model remains close to the reference policy model π_{ref}, via KL divergence between π_θ and π_{ref}.

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. For more introduction of the general version of GRPO, please refer to appendix.

3.2. DocThinker

As illustrated in Fig. 2, we introduce DocThinker, a reinforcement learning (RL)-based framework designed for inference-time reasoning in multimodal document understanding. Unlike conventional supervised fine-tuning (SFT), which is often prone to overfitting and limited generalization, DocThinker refines its reasoning strategies by leveraging rule-based reward signals, enabling greater robustness across diverse document types.

Built on Group Relative Policy Optimization (GRPO), DocThinker optimizes for explainability, adaptability, and accuracy, allowing the model to iteratively improve its decision-making process. While the original GRPO algorithm has been primarily applied to text-only question-answering tasks, DocThinker extends its application to multimodal settings, where both document images and textual queries serve as inputs. Given a document image I and a question q, DocThinker generates a response o_i that includes explainable intermediate reasoning results, such as explicit reasoning traces, rephrased questions for improved clarity, identified Regions of Interest (RoI) supporting the answer, and the final predicted response. Through a multi-objective reward system, DocThinker continuously refines its reasoning strategy via the GRPO algorithm, ensuring stable learning and enhanced adaptability across a wide range of document scenarios. The following sections discuss the choice of base model and prompt template in Sec. 3.2.1, and the design of verifiable multi-objective rewards for document understanding tasks in Sec. 3.2.2.

3.2.1. Base Model and Prompt Template

For our base model, we adopt the state-of-the-art multimodal large language model Qwen2.5-VL 3B and 7B [1], denoted as π_θ, where θ are the trainable parameters. Given a training dataset X, each sample x consists of a document image I and a text prompt p, which includes the user's question q alongside a fixed template message. 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>). The reasoning trace enclosed in <think>...</think>serves as a key component in the model's self-improvement and optimization process during reinforcement fine-tuning. The answer enclosed in <answer>...</answer>is formatted in JSON, containing three critical fields: "rephrase_question", "bbox_2d", and "final_answer". The "rephrase_question" represents an improved, more descriptive version of the original query. This refinement reduces ambiguity, helping users better understand how the model processes the question and formulates an inference. By enhancing question clarity, this component significantly contributes to the model's overall explainability. The "bbox_2d" encodes the 2D bounding box coordinates corresponding to the regions of the document image that the model deems highly relevant to answering the question. This visual cue serves as an explainability aid, providing insights

into which parts of the document influence the model's reasoning. The "final_answer" contains the model's predicted response to the given question q, ensuring that all intermediate reasoning steps contribute to a well-supported final decision. By incorporating both textual and visual reasoning elements, DocThinker enhances explainability across two modalities. The "rephrase_question" field refines the linguistic aspect of reasoning, while the "bbox_2d" field introduces a visual grounding mechanism, making the model's decision-making process more transparent. For further details of the prompt template, please refer to appendix.

Optimization. We adopt the GRPO-based RL objective formulated in Eq. (3) to optimize π_θ, ensuring that the generated answers are accurate, well-structured, and transparently reasoned. The reasoning trace, enclosed within <think>...</think>, serves as a crucial component for self-learning, allowing the model to iteratively refine its reasoning process. Meanwhile, the reward signal, derived from the correctness of the final answer, guides policy optimization, reinforcing high-quality responses. Through this structured reinforcement learning framework, DocThinker achieves greater explainability and generalization in multimodal document understanding.

3.2.2. Multi-objective Reward Functions

The reward model plays a crucial role in reinforcement learning (RL) by aligning the model's predictions with predefined correctness criteria. While traditional RL approaches often rely on human preference-based reward models[20, 50], recent advancements, such as DeepSeek-R1 [5], have demonstrated that verifiable reward functions can significantly enhance a model's reasoning ability. Inspired by this success, we extend Reinforcement Learning with Verifiable Rewards (RLVR) to the visual document understanding domain by designing a rule-based multi-objective reward function that evaluates both textual reasoning and visual comprehension. This ensures that the model not only produces accurate answers but also generates explainable intermediate steps, improving both generalization and transparency. Our framework evaluates model outputs based on four core criteria: format compliance, final answer accuracy, region of interest (RoI) consistency, and rephrased question quality.

Format Reward. The format reward ensures that the model's output adheres to a structured XML-style schema, enforcing consistency for explainable and machine-parsable outputs. It verifies that the reasoning trace is enclosed in <think>...</think>tags and that the final response in <answer>...</answer>is valid JSON with required key-value pairs. Outputs deviating from this format are penalized to maintain structured and systematic reasoning. Given a model output o, the format reward is defined as:

$$R_{\text{format}} = \begin{cases} 1, & \text{if } o \text{ follows the XML-style schema and} \\ & \text{JSON structure with valid key-value pairs,} \\ 0, & \text{otherwise.} \end{cases}$$

Accuracy Reward. The final answer accuracy reward measures whether the model's generated response "final_answer" aligns with the ground truth answer. Unlike traditional RLHF, where correctness is determined through human preference rankings, we leverage a direct verification function that compares the model's final answer against predefined ground truth values. For a given question q and model-generated answer "final_answer", the reward function is defined as:

$$R_{\text{accuracy}} = \begin{cases} 1, & \text{if "final\_answer" = the ground truth,} \\ 0, & \text{otherwise.} \end{cases}$$

RoI IoU Reward. The RoI IoU reward evaluates how accurately the model identifies key visual regions in a document. For a given predicted bounding box B_{pred} = "bbox_2d" and ground truth bounding box B_{gt}, the corresponding reward function is defined as:

$$R_{\text{RoI}} = \begin{cases} 1, & \text{if IoU}(B_{\text{pred}}, B_{\text{gt}}) \geq 0.5, \\ 0, & \text{otherwise.} \end{cases}$$

This reward encourages precise localization of relevant document regions, ensuring extracted information directly supports the model's final answer.

Rephrase Question Reward. To enhance explainability, we introduce a rephrase question reward, which evaluates how effectively the model reframes the original query for improved clarity. Since document-based questions are often ambiguous or underspecified, an ideal model should generate a well-structured and informative rephrased question that provides additional context without altering the intent.

The quality of the rephrased question is assessed based on two criteria: semantic similarity to the original query and word diversity. Given the original question $q_{\text{orig}} = q$ and the rephrased version q_{rephrase} = "rephrase_question", we compute the soft reward, including cosine similarity s and the ratio r of new words compared to the original question:

$$R_{\text{rephrase}} = \begin{cases} s + r, & \text{if } R_{\text{accuracy}} = 1, \\ 0, & \text{otherwise,} \end{cases}$$

where R_{rephrase} is normalized to [0, 1]. This reward encourages the model to generate refined queries that clarify ambiguous input while preserving the original meaning, ultimately improving transparency in its reasoning process.

MLLM	Res.	Data	Str.	Document-oriented Understanding						General Multimodal Understanding				
				Doc/Text					Chart	General VQA		Relation Reasoning		
				DocVQA	TextCaps	TextVQA	DUDE	SROIE	InfoQA	F30k	V7W	GQA	OI	VSR
LLaVA-1.5-7B [22]	336^2	-	SFT	0.244	0.597	0.588	0.290	0.136	0.400	0.581	0.575	0.534	0.412	0.572
LLaVA-1.5-13B [22]	336^2	-	SFT	0.268	0.615	0.617	0.287	0.164	0.426	0.620	0.580	0.571	0.413	0.590
SPHINX-13B [18]	224^2	-	SFT	0.198	0.551	0.532	0.000	0.071	0.352	0.607	0.558	0.584	0.467	0.613
VisCoT-7B [35]	224^2	438k	SFT	0.355	0.610	0.719	0.279	0.341	0.356	0.671	0.580	0.616	**0.833**	0.682
VisCoT-7B [35]	336^2	438k	SFT	0.476	0.675	0.775	0.386	0.470	0.324	0.668	0.558	0.631	0.822	0.614
Qwen2.5VL-7B† [1]	336^2	-	-	0.350	0.642	0.735	0.202	0.472	0.325	0.603	0.556	0.455	0.347	0.616
Qwen2.5VL-7B† [1]	1536^2	-	-	0.773	0.710	0.792	0.492	0.708	0.663	0.685	0.604	0.457	0.371	0.603
Qwen2.5VL-7B* [1]	336^2	4k	SFT	0.355	0.658	0.740	0.215	0.489	0.334	0.624	0.563	0.467	0.405	0.619
Qwen2.5VL-7B* [1]	1536^2	4k	SFT	0.784	0.725	0.801	0.498	0.714	0.674	0.680	0.609	0.472	0.427	0.624
DocThinker-3B	336^2	4k	RL	0.460	0.663	0.746	0.213	0.486	0.335	0.664	0.572	0.486	0.485	0.625
DocThinker-3B	1536^2	4k	RL	0.751	0.691	0.762	0.469	0.735	0.566	0.682	0.583	0.490	0.517	0.637
DocThinker-7B	336^2	4k	RL	0.579	0.682	0.802	0.408	0.495	0.347	0.674	0.580	0.546	0.542	0.656
DocThinker-7B	1536^2	4k	RL	0.795	0.738	0.827	0.515	0.806	0.689	0.701	0.625	0.694	0.686	0.721
DocThinker-7B	1536^2	8k	RL	**0.802**	**0.757**	**0.836**	**0.568**	**0.814**	**0.697**	**0.734**	**0.641**	**0.737**	0.784	**0.768**

Table 1. **Performance on the Visual CoT benchmark.** Grey results indicate zero-shot performance. The final row uses 8k data including F30k, GQA, OI, and VSR; only DUDE, SROIE, V7W are zero-shot. Res. and Str. short for resolution and strategy. InfoQA, F30k, V7W, and OI short for InfographicsVQA, Flickr30k, Visual7W, and Open Images, respectively. † indicate evaluating the model using the official checkpoint. * means trained it on 4data4k setting via supervised fine-tuning. DocThinker and Qwen2.5VL* differ only in training strategy.

Final Reward Function. The total reward combines four rewards to optimize both accuracy and explainability:

$$R_{\text{total}} = \lambda_1 R_{\text{format}} + \lambda_2 R_{\text{accuracy}} + \lambda_3 R_{\text{RoI}} + \lambda_4 R_{\text{rephrase}},$$

where hyperparameters $\lambda_i = 1$ balance reward contributions and avoid reward hacking. This joint optimization ensures the model generates structured, explainable, and verifiable reasoning outputs for document understanding.

4. Experiments

Datasets. We utilize the Visual CoT dataset [35] as training data, which contains 438k question-answer pairs annotated with intermediate bounding boxes that highlight key regions essential for answering questions. These bounding box annotations facilitate the computation of the RoI IoU reward during reinforcement learning, improving the model's ability to focus on relevant areas within document images. The dataset spans five domains, including text/document understanding, fine-grained understanding, charts, general visual question answering (VQA), and relational reasoning.

We establish two training configurations to examine the model's adaptability to different levels of data diversity. The 4data4k setup focuses on document understanding, selecting 1,000 samples each from DocVQA [27], InfographicsVQA [28], TextCaps [37], and TextVQA [38] of Visual CoT dataset, totaling 4,000 instances. The 8data8k configuration extends training data to general VQA and relational reasoning by adding Flickr30k [33], GQA [12], Open Images [15], and VSR [21], with 1,000 samples per dataset, totaling 8,000 instances. This comparison examines how broader domain coverage impacts generalization. Unless specified, we adopt 4data4k configuration as default setting.

Implementation Details. Our base model is Qwen2.5-VL 3B and 7B [1], a state-of-the-art MLLM pretrained on curated web pages, open-source datasets, and synthetic data. To adapt it for document understanding tasks, we train it using the GRPO reinforcement learning framework, as described in Sec. 3. The model is trained for two epochs on eight NVIDIA A100 80GB GPUs, with a batch size of 2. The number of generated candidate responses for GRPO is set to $G = 6$. We employ the AdamW optimizer [25], using a learning rate of $1e-6$. KL coefficient β set to 0.04.

Evaluation. We adopt Visual CoT Benchmark [35], a comprehensive multimodal data, to measure performance across a broad range of document reasoning tasks. Following Visual CoT evaluation protocol, we also assess model's zero-shot ability using SROIE [11], DUDE [16], and Visual7W [54] datasets, evaluating its capacity to generalize beyond its training distribution. We leverage the standard metrics provided by Visual CoT Benchmark [35] to evaluate the model.

4.1. Main Results

Tab. 1 presents the results of our model DocThinker compared to several state-of-the-art MLLMs on the Visual CoT benchmark. The evaluation spans both document-oriented understanding tasks (including text-based document comprehension and chart analysis) and general multimodal understanding (covering general VQA and relational reasoning). Our model demonstrates significant improvements over baseline models, particularly in document understanding tasks, and achieves strong generalization across multiple reasoning domains. Across document-oriented tasks, DocThinker-7B (1536^2, 8k) achieves the highest overall scores, outperforming prior methods, including VisCoT-7B, Qwen2.5VL, and LLaVA variants. The document understanding improvements are largely attributed to the GRPO-based RL strategy with the multi-objective rewards, which enhances the

model's ability to focus on task-relevant regions within documents. In chart understanding (InfoQA), DocThinker-7B (1536^2, 4k) achieves 0.689, surpassing both VisCoT-7B (336^2, 438k, row 5) and Qwen2.5VL-7B (1536^2, row 9), which score 0.324 and 0.674, respectively. This suggests that RL enables better multimodal reasoning, allowing the model to extract and interpret structured information from chart data representations.

Zero Shot Capabilities. As shown in Tab. 1, highlighted results in gray indicate zero-shot generalization performance, where training data splits were not included in the training phase. Our model achieves competitive zero-shot results, particularly in DUDE and SROIE requiring fine-grained text recognition and layout-aware reasoning. Compared to Qwen2.5VL-7B (1536^2, row 7), which achieves 0.492 on DUDE and 0.708 on SROIE, our model reaches 0.568 on DUDE and 0.814 on SROIE. DocThinker outperforms all non-RL-based models, demonstrating that RL with verifiable rewards substantially improves performance in text-heavy, document-based tasks. The performance gap stems from training scale (4k vs. 438k) and zero-shot tasks compared to VisCoT. Still, DocThinker-7B (336^2) outperforms VisCoT-7B (336^2) on all non-zero-shot tasks using only 4k samples and same input. For general multimodal understanding, including VQA and relational reasoning, DocThinker (row 13) continues to achieve competitive performance compared to baselines on Flickr30k (0.701), Visual7W (0.625), GQA (0.694), and VSR (0.721). Particularly in relational reasoning tasks, which require understanding object interactions and contextual relationships, our RL framework enables superior performance. The zero-shot improvements suggest that RL enhances adaptability and generalization, allowing the model to generalize beyond its training distribution and reason effectively in previously unseen scenarios.

RL *vs.* SFT. A direct comparison between supervised fine-tuning (SFT) and RL-based models demonstrates the clear advantage of reinforcement learning. Comparing Qwen2.5VL-7B (SFT, 336^2, 4k, row 8) with DocThinker-7B (RL, 336^2, 4k, row 12), we observe substantial gains across multiple tasks. For example, in DocQA, our model improves from 0.355 to 0.579, and in TextQA, it rises from 0.740 to 0.802. Similarly, in Visual7W, our RL model achieves 0.580 compared to 0.563 with SFT, and in GQA, it improves from 0.467 to 0.546. These improvements highlight that GRPO-based RL training enhances model reasoning and decision-making, enabling it to produce more explainable, structured, and accurate outputs. The multi-objective reward functions, particularly those focused on RoI IoU and rephrased question quality, contribute significantly to the model's ability to focus on relevant document regions and generate clearer reasoning traces.

Explainability Ability. A key advantage of DocThinker over existing MLLMs for document understanding is its ability to generate explicit, explainable reasoning steps, rather than merely providing direct answers. Through our reinforcement learning framework, the model systematically breaks down its thought process into structured intermediate steps, with detailed reasoning enclosed in <think></think>tags and the final response presented within <answer></answer>tags. This structured output enhances transparency, making it easier to analyze how the model arrives at its conclusions. As qualitative performance is shown in Fig. 3, by incorporating structured reasoning rewards, including rephrased question clarity and RoI localization accuracy, our approach significantly enhances explainability over traditional SFT-based MLLMs, ensuring that responses are not only accurate but also more explainable and systematically derived.

Visual Grounding. We further evaluate DocThinker on the TextREC dataset [9], which requires locating objects based on text-conditioned referring expressions. By leveraging RoI IoU reward, DocThinker effectively aligns textual and visual information, improving grounding accuracy. As shown in Tab. 2, DocThinker-7B achieves 82.4% Precision@1, surpassing specialized models like TAMN (80.8%) and MDETR (63.3%). The improvement highlights the impact of reinforcement learning with verifiable rewards, which enhances the model's ability to precisely localize objects tied to scene text. These results demonstrate that DocThinker extends beyond document understanding, excelling in spatial reasoning and multimodal grounding.

Model	Template1	Template2
Specialist Models		
TransVG [6]	50.1	54.0
MAttNet [48]	52.3	60.5
QRNet [45]	52.7	59.1
MDETR [14]	54.4	63.3
TAMN [9]	77.8	80.8
DocThinker-7B	82.4	

Table 2. **Performance on the TextREC dataset.** Precision@1 (%) is reported. Template1 and Template2 represent using two different ways to construct referring expressions proposed in [9]. Template1 is "The object with <OCR string > on it", and Template2 is "The <category name > with <OCR string > on it".

4.2. Ablation Study

Ablating Reward Functions. To assess the impact of reward functions, we ablate RoI IoU (RI) and Rephrase Question (RQ) rewards, analyzing their effect across four datasets, as provided in Tab. 3. Removing RoI IoU reduces performance, particularly on InfoQA and TextCaps, indicating its role in document understanding tasks requiring precise visual grounding. Eliminating Rephrase Question leads to notable

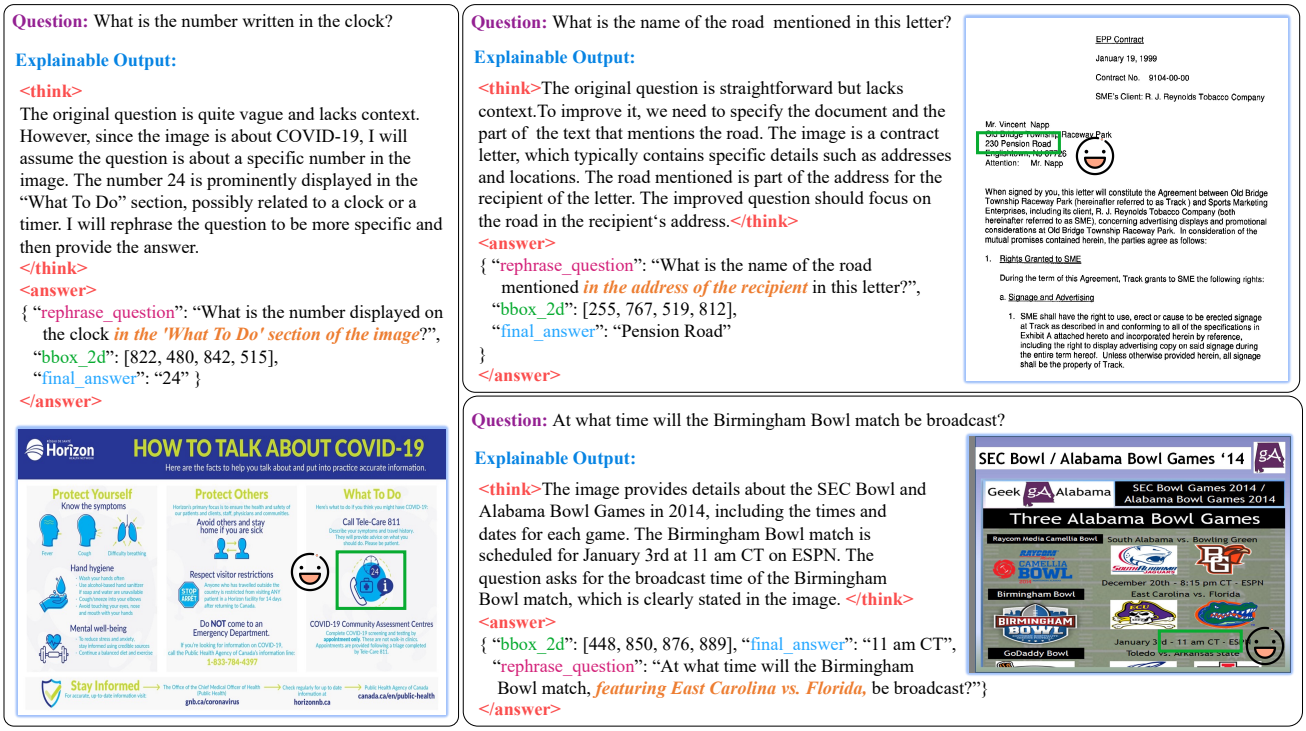

Figure 3. **Qualitative results of DocThinker.** The thinking process significantly improves the reasoning ability and explainability.

drops in TextVQA and DocVQA, highlighting its importance for clarifying ambiguous queries to help the model generate more accurate and explainable responses. The most severe degradation occurs when both are removed, confirming their complementary contributions. This confirms that reinforcement learning with multi-objective rewards is crucial for enhancing both accuracy and reasoning quality in multimodal document understanding.

Method	DocVQA	TextCaps	TextVQA	InfoQA
DocThinker-7B	0.795	0.738	0.827	0.689
w/o RoI IoU	0.775	0.693	0.803	0.637
w/o Rephrase Question	0.763	0.716	0.772	0.658
w/o RI & RQ	0.741	0.662	0.758	0.602

Table 3. **Ablation study of reward functions.** The DocThinker uses 1536^2 input size with 4data4k training data setting. RI and RQ short for RoI IoU and rephrase question reward, respectively.

Ablating KL Divergence. To evaluate the impact of KL divergence regularization, we analyze different values of the coefficient β and its effect on performance across four document-oriented datasets. As shown in Tab. 4, when KL regularization is entirely removed ($\beta = 0$), the model exhibits a decline in performance across all datasets, particularly in TextVQA and DocVQA, indicating that KL regularization plays a key role in stabilizing training and preventing catastrophic forgetting during reinforcement learning. Introducing a small KL weight ($\beta = 0.001$) improves performance slightly but does not fully match the stability compared to setting $\beta = 0.04$. The results suggest that an appropriate KL divergence coefficient is crucial for maintaining balance between exploration and stability, ensuring robust performance across datasets.

Method	DocVQA	TextCaps	TextVQA	InfoQA
DocThinker-7B ($\beta = 0.04$)	0.795	0.738	0.827	0.689
w/o KL ($\beta = 0$)	0.780	0.719	0.803	0.676
$\beta = 0.001$	0.785	0.726	0.812	0.682

Table 4. **Ablation study of the effect of KL Divergence.** The DocThinker uses 1536^2 input size with 4data4k training data setting.

5. Conclusions and Future Works

In this work, we introduced DocThinker, a reinforcement learning-based framework designed to enhance explainability, adaptability, and reasoning ability in multimodal document understanding. By leveraging Group Relative Policy Optimization and a multi-objective reward system, our approach dynamically refines reasoning strategies at inference time, overcoming the limitations of static Chain-of-Thought reasoning and supervised fine-tuning. DocThinker achieves state-of-the-art or highly competitive performance on standard benchmarks compared with previous SFT-based methods across multiple document understanding tasks. Expanding DocThinker to larger multimodal models and broader reasoning tasks and scenarios, such as scientific and legal document analysis, could further improve adaptability.

Acknowledgements. This work was supported by the NSFC (62225603, 62206104), and Alibaba AIR program.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 4, 6

[2] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. 2

[3] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *ArXiv*, abs/2501.17161, 2025. 3

[4] Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie Zhang, Chen Qian, Kun Wang, et al. Explainable and interpretable multimodal large language models: A comprehensive survey. *ArXiv*, arXiv:2412.02104, 2024. 1, 3

[5] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bing-Li Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dong-Li Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Jiong Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, M. Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shao-Kang Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wen-Xia Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyu Jin, Xi-Cheng Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yi Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yu-Jing Zou, Yujia He, Yunfan Xiong, Yu-Wei Luo, Yu mei You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yao Li, Yi Zheng, Yuchen Zhu, Yunxiang Ma, Ying Tang, Yukun Zha, Yuting Yan, Zehui Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhen guo Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zi-An Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *ArXiv*, abs/2501.12948, 2025. 2, 3, 5

[6] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wen gang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *IEEE/CVF International Conference on Computer Vision*, pages 1749–1759, 2021. 7

[7] 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. 2

[8] Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei A. F. Florêncio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. *ArXiv*, abs/2501.05452, 2025. 1, 3

[9] Chenyu Gao, Biao Yang, Hao Wang, Mingkun Yang, Wenwen Yu, Yuliang Liu, and Xiang Bai. Textrec: A dataset for referring expression comprehension with reading comprehension. In *IEEE International Conference on Document Analysis and Recognition*, 2023. 7

[10] Weiche Hsieh, Ziqian Bi, Chuanqi Jiang, Junyu Liu, Benji Peng, Sen Zhang, Xuanhe Pan, Jiawei Xu, Jinlang Wang, Keyu Chen, Pohsun Feng, Yizhu Wen, Xinyuan Song, Tianyang Wang, Ming Liu, Junjie Yang, Ming Li, Bowen Jing, Jintao Ren, Jun-Jie Song, Hong-Ming Tseng, Yichao Zhang, Lawrence K.Q. Yan, Qian Niu, Silin Chen, Yunze Wang, and Chia Xin Liang. A comprehensive guide to explainable ai: From classical models to llms. *ArXiv*, abs/2412.00800, 2024. 1

[11] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *International Conference on Document Analysis and Recognition*, pages 1516–1520, 2019. 6

[12] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6693–6702, 2019. 6

[13] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. 3

[14] Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. Mdetr - modulated detection for end-to-end multi-modal understanding. In *IEEE/CVF International Conference on Computer Vision*, pages 1760–1770, 2021. 7

[15] Alina Kuznetsova, Hassan Rom, Neil Gordon Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4. *International Journal of Computer Vision*, 128:1956 – 1981, 2018. 6

[16] Jordy Van Landeghem, Rubèn Pérez Tito, Łukasz Borchmann, Michal Pietruszka, Pawel J'oziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Ackaert, Ernest Valveny, Matthew B. Blaschko, Sien Moens, and Tomasz Stanislawek. Document understanding dataset and evaluation (dude). In *IEEE/CVF International Conference on Computer Vision*, pages 19471–19483, 2023. 6

[17] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vuli'c, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *ArXiv*, abs/2501.07542, 2025. 2

[18] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Jiao Qiao. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *ArXiv*, abs/2311.07575, 2023. 6

[19] Chenglong Liu, Haoran Wei, Jinyue Chen, Lingyu Kong, Zheng Ge, Zining Zhu, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Focus anywhere for fine-grained multi-page document understanding. *arXiv preprint arXiv:2405.14295*, 2024. 2

[20] Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. Skywork-Reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*, 2024. 5

[21] Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2022. 6

[22] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26286–26296, 2023. 6

[23] 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. 2

[24] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *ArXiv*, 2025. 2, 3

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 6

[26] Pengyuan Lyu, Yulin Li, Hao Zhou, Weihong Ma, Xingyu Wan, Qunyi Xie, Liang Wu, Chengquan Zhang, Kun Yao, Errui Ding, and Jingdong Wang. Structextv3: An efficient vision-language model for text-rich image perception, comprehension, and beyond. *ArXiv*, abs/2405.21013, 2024. 2

[27] Minesh Mathew, Dimosthenis Karatzas, R. Manmatha, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In *IEEE Winter Conference on Applications of Computer Vision*, pages 2199–2208, 2020. 6

[28] Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. Infographicvqa. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2582–2591, 2021. 6

[29] OpenAI. ChatGPT. https://openai.com/chatgpt, 2023. Accessed: 2025-02-20. 1

[30] OpenAI. GPT-4. https://openai.com/gpt-4, 2023. Accessed: 2025-02-25.

[31] OpenAI. GPT-4V(ision) System Card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023. Accessed: 2025-02-26. 1

[32] Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Ouyang Cheng, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *ArXiv*, 2025. 2, 3

[33] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, J. Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123:74–93, 2015. 6

[34] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017. 3

[35] 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, 6

[36] 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. 2, 3

[37] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, pages 742–758, 2020. 6

[38] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8309–8318, 2019. 6

[39] Shilin Sun, Wenbin An, Feng Tian, Fang Nan, Qidong Liu, Jun Liu, Nazaraf Shah, and Ping Chen. A review of multimodal explainable artificial intelligence: Past, present and future. *ArXiv*, abs/2412.14056, 2024. 1, 3

[40] Richard S. Sutton and Andrew G. Barto. Introduction to reinforcement learning. *ArXiv*, 1998. 2

[41] Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. Docllm: A layout-aware generative language model for multimodal document understanding. In *Annual Meeting of the Association for Computational Linguistics*, 2023. 2

[42] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language model. In *European Conference on Computer Vision*, pages 408–424. Springer, 2024. 2

[43] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In *Neural Information Processing Systems*, pages 24824–24837, 2022. 1

[44] Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind's eye of llms: Visualization-of-thought elicits spatial reasoning in large language models. In *Neural Information Processing Systems*, 2024. 2

[45] Jiabo Ye, Junfeng Tian, Ming Yan, Xiaoshan Yang, Xuwu Wang, Ji Zhang, Liang He, and Xin Lin. Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15481–15491, 2022. 7

[46] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Mingshi Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, Qiang Qi, Ji Zhang, and Feiyan Huang. mplug-docowl: Modularized multimodal large language model for document understanding. *ArXiv*, abs/2307.02499, 2023. 2

[47] 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. *arXiv preprint arXiv:2310.05126*, 2023. 2

[48] Licheng Yu, Zhe L. Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. Mattnet: Modular attention network for referring expression comprehension. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018. 7

[49] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024. 3

[50] Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, et al. A simple yet effective multi-modal reward model. *arXiv preprint arXiv:2501.12368*, 2025. 5

[51] Jinxu Zhang. Read and think: An efficient step-wise multimodal language model for document understanding and reasoning. *ArXiv*, 2024. 3

[52] Jiaxin Zhang, Wentao Yang, Songxuan Lai, Zecheng Xie, and Lianwen Jin. Dockylin: A large multimodal model for visual document understanding with efficient visual slimming. *ArXiv*, abs/2406.19101, 2024. 2

[53] 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. 2

[54] Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4995–5004, 2015. 6

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.

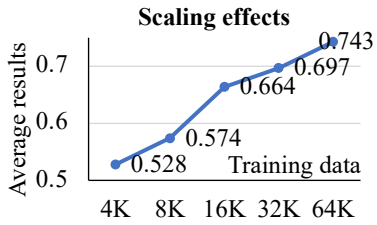

Figure 1. Scaling effects of DocThinker on document benchmarks. Increasing training data from 4k to 64k consistently improves performance, showing the scalability of GRPO.

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.

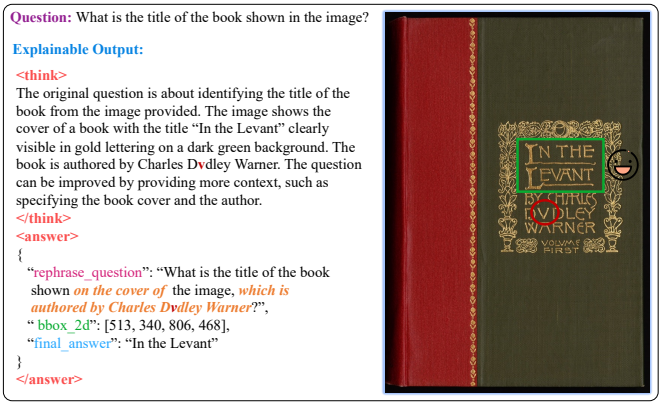

Figure 2. Failure case. Although the model correctly predicts the final answer "In the Levant" and localizes the 2D bounding box of RoI accurately, the rephrased question contains a hallucination: it misidentifies the author as "Charles Dvdley Warner" instead of "Charles Dudley Warner". This error likely 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