RegionDoc-R1: Reinforcing Semantic Layout-Aware Reasoning for Document Understanding

Anonymous ACL submission

Abstract

Reinforcement Learning (RL) has proven effective in eliciting reasoning abilities of Multimodal Large Language Models (MLLMs) and is emerging in document understanding. However, existing approaches usually lack fine-grained region-level grounding supervision, while the key challenge in document reasoning lies in linking visual evidence within precise regions directly to question–answer semantics. In this work, we introduce ***RegionDoc-R1***, a novel framework for document understanding that enhances MLLMs' reasoning with step-wise feedback. Instead of previous OCR-based text positions, we encourage the model to predict question-relevant Regions of Interest (RoIs) and ground the answer in these regions. Specifically, we first propose the Region-Aware Group Relative Policy Optimization (*RA-GRPO*), which encourages models to utilize region-level spatial information in documents for reasoning. Correspondingly, we construct a hybrid training corpus, named *SR-Doc*, containing Semantic Reasoning (SR) examples enriched with cross-page and region-level reasoning layout annotations. Meanwhile, we also introduce an Adaptive Chain-of-Thought (*Ada-CoT*) strategy, which dynamically adjusts the reasoning process according to different tasks, enabling more efficient and flexible step-wise document understanding. Experiments on several document reasoning benchmarks demonstrate that RegionDoc-R1 achieves state-of-the-art performance across tasks such as form document understanding, table-based QA, and layout-sensitive information extraction.[1]

1 Introduction

Document understanding is a fundamental task at the intersection of Natural Language Processing (NLP) and computer vision, with applications in information retrieval (Rajpurkar et al., 2016; Yang et al., 2018), question answering (Hermann et al., 2015; Chen et al., 2017), enterprise knowledge management (Li et al., 2020; Gu et al., 2021), information extraction from Visually Rich Documents (Perot et al., 2024; Bhattacharyya et al., 2025), and legal or financial analysis (Chalkidis et al., 2020; Kornilova and Eidelman, 2019). Early approaches typically relied on Optical Character Recognition (OCR) to extract plain text from documents, which was then processed using standard NLP pipelines. However, OCR-only methods struggle to capture the rich multimodal nature of real-world documents, which often contain images, tables, and other structured visual elements. To overcome this limitation, researchers proposed treating entire documents as images and applying computer vision techniques, such as convolutional neural networks and vision transformers, to capture holistic visual cues (Appalaraju et al., 2021; Garncarek et al., 2021; Xu et al., 2021). These methods can effectively leverage visual signals such as font styles, graphical elements, or color usage. Nevertheless, they often fail to handle densely packed text regions or complex structured components like tables and figures, which demand fine-grained reasoning over both visual and textual information.

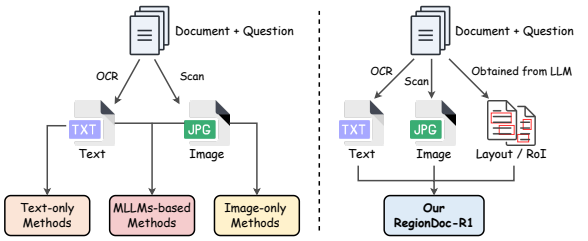


Figure 1: Comparison of different paradigms in document understanding. Left: Traditional pipelines obtain plain text via OCR or images via scanning, which are then processed by text-only, image-only, or MLLMs-based methods. Right: RegionDoc-R1 predicts question-relevant RoIs as part of the model output for evidence grounding, enabling region-level reasoning across pages (RoIs are not required as inputs during inference).

[1] Code and dataset will be public in github.

With the rise of Multimodal Large Language Models (MLLMs) (Liu et al., 2023; Bai et al., 2023), recent studies jointly process textual and visual modalities for document understanding (Ye et al., 2023b; Zhang et al., 2023). For instance, AlignVLM (Masry et al., 2026) focuses on improving vision-language alignment to achieve stronger performance across diverse tasks, yet it overlooks the *layout structure*, which often leads to reasoning failures for document-level tasks, particularly when information must be read in a structured order rather than a linear sequence. Document layout plays a crucial role in conveying semantic relationships between elements such as headers, footnotes, columns, or table cells (Xu et al., 2020; Powalski et al., 2021). Beyond modeling such relationships, layout information also enables *region-specific reasoning*, where only a localized portion of the document is relevant to a given question or task. This mirrors human reading strategies that readers typically locate the pertinent region before focusing on deeper understanding (Lee et al., 2022; Garncarek et al., 2021), as shown in Figure 1.

In parallel, Reinforcement Learning (RL) has recently emerged as an effective paradigm for enhancing reasoning in LLMs. For example, DeepSeek-R1 (Guo et al., 2025) demonstrates that large-scale RL can significantly boost Chain-of-Thought (CoT) reasoning and verification abilities. Beyond text-based reasoning, RL has also been applied in multimodal scenarios such as video understanding, where Video-R1 (Feng et al., 2026) leverages preference optimization to improve temporal reasoning and action comprehension. These advances suggest that RL offers a powerful mechanism to align MLLMs toward more accurate and human-like reasoning. Recently, in the domain of document understanding, DocThinker (Yu et al., 2025) employs rule-based strategies to improve explainability, while DocR1 (Xiong et al., 2026) applies GRPO to enable multi-page reasoning. However, both approaches remain limited: DocThinker mainly assumes single-page evidence, and DocR1 only provides page-level grounding without precise Regions of Interest (RoI) localization.

To address these limitations, we introduce RegionDoc-R1, a region-aware reinforcement learning framework for semantic document reasoning. Unlike previous efforts, RegionDoc-R1 introduces *Region-Aware Group Relative Policy Optimization (RA-GRPO)*, which leverages region-level semantic grounding to precisely localize fine-grained ROIs, even when they span multiple pages and are directly linked to question-answer semantics. This region-level perspective enables more accurate reasoning and better reflects human strategies of first identifying relevant areas and then performing focused understanding. To support this, we curate a new hybrid corpus, *SR-Doc*, which incorporates high-quality Semantic Reasoning (SR) layouts obtained from LLMs, and provides diverse instances with explicit multi-page and ROI-oriented annotations. In addition, we design an *Adaptive Chain-of-Thought (Ada-CoT)* mechanism that tailors the reasoning trajectory to different questions, thereby improving both precision and contextual awareness during reasoning. Together, these innovations equip RegionDoc-R1 with the ability to integrate region-level grounding and adaptive reasoning, thus advancing the state-of-the-art in various document understanding tasks.

Our main contributions include:

- We introduce ***RegionDoc-R1*** as a new reinforcement learning framework for *region-level, layout-aware* document reasoning within MLLMs, paving the way for adaptable and robust document understanding.

- We propose a Region-Aware Group Relative Policy Optimization (*RA-GRPO*) method that exploits semantic layouts for reasoning, supported by a new hybrid dataset *SR-Doc* with cross-page RoI-structured annotations, and complemented by an Adaptive Chain-of-Thought (Ada-CoT) strategy for dynamic question-specific reasoning.

- We conduct extensive experiments on multiple benchmarks, showing that RegionDoc-R1 outperforms existing methods and demonstrates the benefits of adaptive ROI-based RL for multi-page document understanding

2 Related Work

2.1 Multimodal Document Understanding

Large Language Models (LLMs) and Multimodal LLMs (MLLMs) have shown strong capabilities in understanding visually rich documents by leveraging both textual and layout information. Traditional methods such as LayoutLM series (Xu et al., 2020, 2021; Huang et al., 2022) incorporate layout and image features to improve performance in tasks such as form understanding and receipt parsing. Recently, instruction tuning models such as

LLaVA-Doc (Liu et al., 2023) further adapt LLMs to document understanding using multimodal data. mPLUG-DocOwl (Ye et al., 2023a) incorporates multi-task instruction tuning over diverse document understanding datasets, including captioning, VQA, and information extraction. Unified models like UniDoc (Gu et al., 2021) and UDOP (Tang et al., 2023) generalize across various document tasks via a sequence-to-sequence formulation, enabling zero-shot and few-shot reasoning. More recently, LayoutLLM (Luo et al., 2024) introduces a layout-aware tuning strategy that integrates document pre-trained models with fine-tuned LLMs. However, while prior research focuses on supervised pre-training or fine-tuning, document reasoning MLLMs through RL remains under-explored.

2.2 Large Language Model Reasoning

Reasoning with LLMs has attracted significant research interest due to their emerging capabilities in multi-step problem-solving. A key line of work focuses on *Chain-of-Thought (CoT)* prompting, which encourages models to explicitly decompose complex tasks into intermediate steps (Wei et al., 2022; Kojima et al., 2022). This strategy is shown to significantly improve performance in arithmetic, commonsense, and symbolic reasoning benchmarks. To scale up CoT data, automatic generation methods have been proposed (Wang et al., 2023; Zhou et al., 2023), often relying on strong teacher models, such as GPT-4 (Achiam et al., 2023). For example, LLaVA-CoT (Xu et al., 2025) employs structured CoT templates (*e.g.*, summary, caption, reasoning, conclusion) for SFT, while Mulberry (Yao et al., 2026) proposes CoMCTS, a collaborative Monte Carlo Tree Search method for discovering diverse reasoning paths. Some studies investigate self-improvement via Reinforcement Learning (RL). DeepSeek-R1 (Guo et al., 2025) shows that even coarse, outcome-only rewards can guide models to generate effective CoT reasoning via rule-based RL. Follow-up research, such as Open Reasoner Zero (Hu et al., 2026) and Kimi-1.5 (Team et al., 2025), further validates this paradigm in both text and vision domains. Extending RL to multimodal document understanding is a natural next step, especially when reasoning requires both semantic and structural cues. However, effectively exploiting layout information remains challenging.

2.3 Layout Modeling in Document Understanding

Modeling the spatial layout is crucial for understanding structured information in visually rich documents. Early studies such as LayoutLM (Xu et al., 2020), LayoutLMv2 (Xu et al., 2021), and LayoutLMv3 (Huang et al., 2022) jointly encode textual content and 2D positions, while extensions like StrucTexT (Li et al., 2021) and XYLayoutLM (Gu et al., 2022) capture fine-grained spatial and hierarchical structures. OCR-free methods (*e.g.*, Donut (Kim et al., 2022) and TILT (Powalski et al., 2021)) instead learn layout implicitly from raw pixels, though they struggle with explicit reasoning in multi-page or form-like documents. More recently, tuning-based multimodal models (*e.g.*, LLaVAR (Zhang et al., 2023), LayoutLLM (Luo et al., 2024)) incorporate layout cues indirectly, *e.g.*, through OCR-augmented captions or coarse layout embeddings. DocThinker (Yu et al., 2025) introduces rule-based reinforcement learning for explainable reasoning but assumes evidence within a single page, while DocR1 (Xiong et al., 2026) applies GRPO to multi-page settings yet operates at a coarse page level without fine-grained region grounding. In contrast, our RegionDoc-R1 introduces a layout-aware reinforcement learning framework that explicitly grounds reasoning in semantically relevant *region-level* RoIs across pages, enabling more accurate and robust document understanding. Unlike these layout modeling methods relying on explicit structural representations, ours goes further by dynamically predicting task-relevant regions during reasoning without requiring predefined layout annotations, bridging spatial structure with semantic reasoning.

3 Methodology

Our goal is to develop a reinforcement learning framework that enhances region-aware reasoning for multimodal document understanding. While supervised fine-tuning (SFT) can endow models with general reasoning ability, it often fails to generalize robustly across diverse document layouts and tasks. To overcome these challenges, we propose RegionDoc-R1, which augments SFT initialization with reinforcement learning guided by multiple reward signals. We explicitly link reasoning steps with semantically relevant Regions of Interest (RoIs) across pages. Beyond grounding, we further introduce an Adaptive Chain-of-Thought

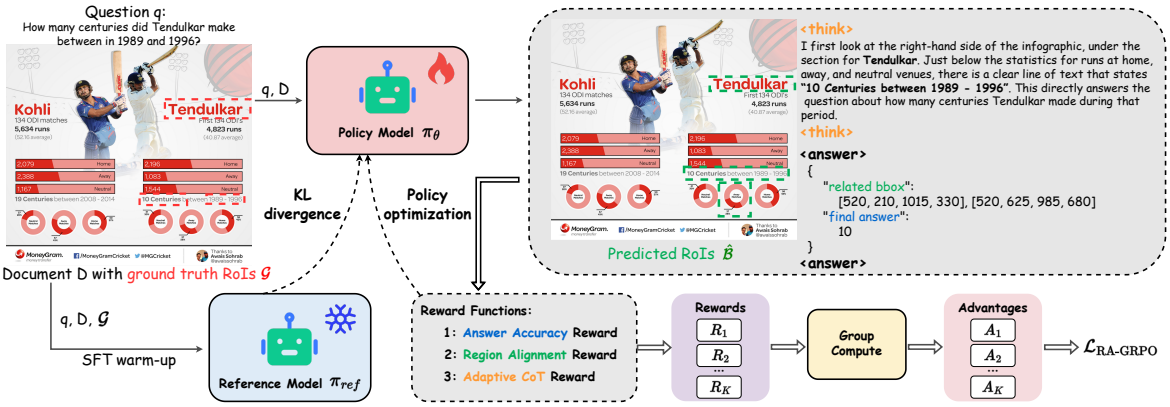


Figure 2: Overview of the proposed ***RegionDoc-R1*** framework. Given a document image D and a question q, we provide ground-truth regions of interest (RoIs) $\mathcal{G}$ for SFT initialization to obtain the reference model π_{ref}. The policy model π_θ then predicts reasoning trajectories including intermediate thoughts, supporting RoIs, and final answers. Each predicted output is evaluated by our multi-objective reward functions (Sec. 3.1), including *Answer Accuracy Reward*, *Region Alignment Reward*, and *Adaptive CoT Reward*. The obtained rewards are aggregated and normalized into group-relative advantages, which are used to optimize the policy model via the $\mathcal{L}_{\text{RA-GRPO}}$ objective, while keeping it aligned with the reference model π_{ref} through KL regularization.

(Ada-CoT) mechanism to dynamically adjust reasoning traces according to different tasks, enabling both efficient and context-sensitive inference.

As shown in Figure 2, our method is based on GRPO (see Appx. A for details). Given a document and a question, the policy model predicts reasoning trajectories consisting of intermediate thoughts, localized RoIs, and final answers. Each candidate output is then evaluated by a set of complementary reward functions: (1) an *Answer Accuracy Reward* that verifies the correctness of final predictions; (2) a *Region Alignment Reward* that enforces precise spatial grounding against annotated RoIs; and (3) an *Adaptive CoT Reward* that regulates reasoning length and routing. The obtained rewards are normalized into group-relative advantages, which drive policy optimization through the objective, while a reference model obtained from supervised fine-tuning serves as a regularization anchor. Note that our model takes only raw document images and textual questions as input, without relying on external OCR systems. Textual information is implicitly extracted by the multimodal backbone, Qwen2.5-VL (Wang et al., 2024), which has been pretrained on large-scale OCR-rich data.

3.1 Region-Aware Group Relative Policy Optimization (RA-GRPO)

Standard GRPO (Shao et al., 2024b) is a critic-free RL framework based on group-relative advantages, yet it overlooks spatial layout cues and assumes fixed-length reasoning traces. We therefore propose Region-Aware GRPO (RA-GRPO), which incorporates region-level layout supervision from *SR-Doc* and an Adaptive Chain-of-Thought (Ada-CoT) mechanism, enabling policies to answer accurately, ground their predictions in semantically relevant RoIs, and dynamically adjust reasoning strategies.

Given a question-document pair (q, D), the policy model π_θ samples K responses $\{(o^{(i)}, \hat{\mathcal{B}}^{(i)})\}_{i=1}^{K}$. Each candidate is scored by a multi-objective reward $R^{(i)}$, and the group-normalized advantage A_i is:

$$A_i = \frac{R^{(i)} - \mu}{\sigma}, \qquad \mu = \frac{1}{K}\sum_{i=1}^{K} R^{(i)},$$
$$\sigma = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(R^{(i)} - \mu)^2 + \epsilon}. \tag{1}$$

Answer Accuracy Reward. First, we check whether the predicted final answer o_{final} matches the ground-truth y. For *textual answers*, we directly check whether the predicted final answer o_{final} matches the ground-truth y after normalization:

$$R_{\text{ans}} = \begin{cases} 1, & \text{if } \text{norm}(o_{\text{final}}) = \text{norm}(y), \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

where $\text{norm}(\cdot)$ normalizes case and spacing. For *numeric answers*, we allow approximate matching by measuring the relative error e with tolerance δ:

$$e = \frac{|o_{\text{final}} - y|}{\max(|y|, 1)}, \qquad R_{\text{ans}} = \exp\left(-\frac{e}{\delta}\right). \tag{3}$$

We adopt normalized exact matching as the reward signal for textual answers due to its simplicity and stability in RL optimization.

Region Alignment Reward. To enforce spatial grounding, predicted RoIs $\hat{\mathcal{B}}$ are compared with the verified ground-truth RoIs $\mathcal{G}$ from SR-Doc. We compute IoU-based precision, recall, and F_1:

$$\text{Prec} = \frac{|\mathcal{M}|}{|\hat{\mathcal{B}}|}, \qquad \text{Rec} = \frac{|\mathcal{M}|}{|\mathcal{G}|},$$
$$F_1 = \frac{2 \cdot \text{Prec} \cdot \text{Rec}}{\text{Prec} + \text{Rec} + \varepsilon}. \tag{4}$$

where $\mathcal{M}$ is the set of matched box pairs under threshold τ. We further measure region coverage:

$$\text{Cov} = \frac{\sum_p \text{Area}(\cup_{m:p_m=p} \hat{b}_m \cap \cup_{j:g_j.p=p} g_j)}{\sum_p \text{Area}(\cup_{j:g_j.p=p} g_j) + \varepsilon}, \tag{5}$$

where $\hat{b}_m$ and g_j denote predicted and ground-truth RoIs on page p, respectively. This metric measures the fraction of ground-truth area covered by predictions, complementing F_1 by emphasizing spatial completeness. The region reward R_{roi} is:

$$R_{\text{roi}} = \alpha_1 F_1 + \alpha_2 \text{Cov} - \alpha_3 \cdot \frac{\max(|\hat{\mathcal{B}}| - |\mathcal{G}|, 0)}{\max(|\hat{\mathcal{B}}|, 1)}, \tag{6}$$

where α_1 and α_2 balance the contributions of F_1 accuracy and coverage, while α_3 controls redundancy in predicted regions, complementing the coverage term by discouraging excessive overlapping or irrelevant regions. This formulation combines region-level accuracy (F1) and global coverage (Cov), which together provide a more comprehensive measure of spatial grounding quality.

Adaptive CoT Reward. We further introduce Ada-CoT to introduce dynamic control over whether and how reasoning traces are generated. The model first samples a routing variable $z \in \{\text{direct}, \text{cot}\}$:

$$p_\theta(z \mid q, D) = \text{softmax}(W^\top h(q, D)), \tag{7}$$

where $h(q, D)$ denotes the joint representation of the input question q and document D, obtained from the hidden states of the underlying MLLM encoder. If $z = \text{direct}$, the model outputs an answer directly; if $z = \text{cot}$, it produces a reasoning trace τ consisting of $|\tau|$ steps.

Thus, we design two reward components:

$$R_{\text{gate}} = \begin{cases} 1, & \text{if } z = \text{direct},\ y = \hat{y}, \\ \kappa, & \text{if } z = \text{cot}, \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$

where $y = \hat{y}$ means the prediction is correct, thus R_{gate} encourages the model to answer directly when possible, while still rewarding CoT routing with a base score κ.

$$R_{\text{len}} = \exp\Big(- \frac{\max(|\tau| - B_0, 0)}{\delta_{\text{len}}} \Big), \tag{9}$$

Where R_{len} softly penalizes traces longer than B_0, while allowing shorter ones to retain full reward.

The overall Ada-CoT reward is defined as:

$$R_{\text{adacot}} = \rho_1 R_{\text{gate}} + \rho_2 R_{\text{len}}. \tag{10}$$

We term this mechanism "adaptive" as the model dynamically selects between direct answering and multi-step reasoning based on the input, rather than following a fixed Chain-of-Thought pattern. This adaptive routing allows the model to avoid unnecessary reasoning for simple queries while preserving multi-step reasoning for complex cases. Importantly, although direct answers receive a higher reward, correctness and region-aware rewards dominate the overall objective. When multi-step reasoning is required, direct answering typically leads to incorrect predictions and thus lower overall reward, naturally encouraging the model to generate reasoning chains when necessary.

Final Reward and Loss. The overall reward is

$$R = \lambda_{\text{ans}} R_{\text{ans}} + \lambda_{\text{roi}} R_{\text{roi}} + \lambda_{\text{adacot}} R_{\text{adacot}}. \tag{11}$$

The RA-GRPO loss is then

$$\mathcal{L}_{\text{RA-GRPO}}(\theta) = -\frac{1}{K} \sum_{i=1}^{K} \frac{1}{|o^{(i)}|} \sum_{t=1}^{|o^{(i)}|} \Big[\rho_{i,t}(\theta) A_i - \beta\, \mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}) \Big], \tag{12}$$

$$\rho_{i,t}(\theta) = \frac{\pi_\theta(o_t^{(i)} \mid q, D, o_{<t}^{(i)})}{\varphi[\pi_\theta(o_t^{(i)} \mid q, D, o_{<t}^{(i)})]}. \tag{13}$$

where A_i is computed as in Eq. 1, and π_{ref} denotes the reference policy (the SFT-initialized checkpoint), serving as an anchor to regularize training via KL divergence. Hyperparameter settings and a detailed pseudo-code of RA-GRPO are in Appx. B.

3.2 SR-Doc Dataset

We construct SR-Doc as a region-aware training corpus for document reasoning. Each QA pair is associated with question-relevant RoIs. The

RoIs are automatically generated by Qwen2-VL-72B-Instruct and filtered through crop-based sufficiency verification: a QA–RoI pair is retained if the cropped RoIs alone can reproduce the ground-truth answer. SR-Doc contains approximately 130k documents and 450k QA pairs, covering both single-page and multi-page scenarios (See Appx. C).

4 Experiments

Training. We adopt a two-stage training strategy leveraging both existing multimodal reasoning resources and our newly constructed dataset. In the SFT stage, we use the Visual CoT dataset (Shao et al., 2024a), which contains 438k question-answer pairs with annotated bounding boxes highlighting key regions required to answer the questions. These annotations span five domains, including text/document understanding, fine-grained recognition, charts, general visual question answering, and relational reasoning. The bounding box supervision equips the model with generic multimodal reasoning skills and the ability to ground predictions in relevant visual regions.

In the RL stage, we rely on our proposed *SR-Doc* dataset that provides multiple region-level RoI annotations over multi-page documents. Unlike Visual CoT, SR-Doc focuses specifically on document understanding tasks, offering direct spatial supervision that aligns with our layout-aware reward design. This enables fine-grained policy optimization and robust reasoning across diverse tasks.

Evaluation. We evaluate our model on a broad suite of document understanding datasets, grouped into *single-page* and *multi-page* settings. For single-page document reasoning, we include seven widely used datasets: DocVQA (Task 1) (Mathew et al., 2021), InfoGraphicVQA (Mathew et al., 2022), TabFact (Chen et al., 2020), TextCaps (Sidorov et al., 2020), TextVQA (Singh et al., 2019), SROIE (Huang et al., 2019), and VisualMRC (Tanaka et al., 2021). These datasets span diverse domains such as forms, charts, tables, and visually-rich natural images, providing a comprehensive evaluation for region-level document reasoning and fine-grained visual-text grounding. To test cross-page reasoning ability, we adopt SlideVQA (Tanaka et al., 2023), ArxivQA (Xiong et al., 2026), DUDE (Van Landeghem et al., 2023), and MP-DocVQA (Mathew et al., 2021). These corpora feature long documents (*e.g.*, scientific articles, enterprise reports) where evidence is distributed across pages, stressing cross-page grounding and layout-aware reasoning. Together, these benchmarks enable a comprehensive assessment of document reasoning across both localized evidence extraction and long-range document understanding.

Implementation Details. Our RegionDoc-R1 is built on Qwen2.5-VL-7B and trained on 8 NVIDIA H100 GPUs with mixed-precision optimization. We adopt a two-stage training strategy: SFT on Visual CoT for 3 epochs, followed by RL on *SR-Doc* for 1.5 epochs. During RL, the maximum input resolution is increased from 896×896 to 1024×1024 to better capture fine-grained RoIs. Training takes approximately 6 days for SFT and 4 days for RL (more details in Appx. D).

4.1 Main Results

We evaluate RegionDoc-R1 against recent state-of-the-art multimodal LLMs on both single-page and multi-page document reasoning benchmarks.

RegionDoc-R1 achieves strong results across all seven single-page benchmarks as shown in Table 1. Compared with LLaVA-1.5-7B (Liu et al., 2024) and VisCoT-7B (Shao et al., 2024a), our approach yields clear gains on layout-intensive tasks such as DocVQA and InfoVQA. Although DocOwl series (Ye et al., 2023a; Hu et al., 2025) and Qwen2.5-VL (Wang et al., 2024) are competitive, they mainly rely on image-text alignment rather than explicit region grounding. Our layout-aware RL provides an advantage on datasets like SROIE and InfoVQA, where precise RoI localization is critical. While DocThinker and DocR1 leverage RL, the former uses rule-based strategies and the latter stays at page-level granularity. By contrast, our fine-grained RoI grounding explains the improvements on TextVQA and SROIE. On TabFact, where reasoning is largely textual, DocOwl-1.5 remains slightly ahead, showing the complementarity of factual reasoning and spatial grounding.

We also observe a slight performance drop on VisualMRC compared to Qwen2.5-VL. We attribute this to the nature of VisualMRC, which places less emphasis on explicit region-level grounding. Our method introduces additional grounding constraints through region-aware rewards, which may slightly limit flexibility in such cases. Nevertheless, our approach consistently improves tasks requiring precise region grounding, demonstrating its effectiveness for document reasoning.

Table 2 shows that RegionDoc-R1 achieves the

Model	Param.	DocVQA	InfoVQA	TabFact	TextCaps	TextVQA	SROIE	VisualMRC
LLaVA-1.5-7B	7B	61.2	34.5	60.7	56.3	54.8	58.1	210.4
VisCoT-7B	7B	69.7	39.8	65.9	58.2	61.7	62.4	221.6
UReader	7B	66.8	43.1	68.9	59.8	58.7	61.2	223.5
TextMonkey	9B	72.4	30.7	65.4	61.0	64.2	63.3	229.1
DocOwl-1.5-Chat	8B	83.1	52.6	**80.8**	62.3	70.1	65.7	242.0
DocOwl-2	8B	81.5	48.2	77.4	60.5	67.5	66.9	219.3
Qwen2.5-VL	7B	94.2	80.9	77.5	64.1	85.1	70.4	**255.8**
DocThinker*	7B	80.2	–	–	**75.7**	83.6	81.4	–
DocR1*	7B	95.1	82.6	79.6	–	81.0	–	251.6
RegionDoc-R1	7B	**95.3**	**83.2**	79.9	75.1	**85.8**	**81.7**	252.4

Table 1: Comparison on seven single-page document reasoning benchmarks. The best results are highlighted in **bold**. Mark * means results are taken directly from the original paper. "–" indicates that the result is not reported. For DocThinker, we report the setting with 1536^2 input resolution and 8K training data (highest reported in paper).

Model	Param.	MP-DocVQA	DUDE	SlideVQA	ArxivQA
LLaVA-1.5-7B	7B	39.12±0.25	25.20±0.18	28.96±0.22	5.40±0.10
VisCoT-7B	7B	49.61±0.28	31.20±0.25	45.52±0.26	10.50±0.14
DocOwl-2	8B	68.15±0.32	32.02±0.21	29.53±0.19	7.19±0.11
Qwen2.5-VL	7B	87.44±0.35	52.85±0.27	70.32±0.33	29.60±0.22
DocR1*	7B	87.45	54.39	71.96	29.57
RegionDoc-R1	7B	**88.32**±0.34	**55.71**±0.29	**73.45**±0.35	**42.57**±0.25

Table 2: Performance comparison on four multi-page document reasoning benchmarks. Mark * means results are taken directly from the original paper. The best results are highlighted in **bold**.

best results across all four multi-page benchmarks. Compared with Qwen2.5-VL, our method yields consistent gains, especially on ArxivQA, where region-level RoI grounding provides more precise supervision for long scientific articles. On MP-DocVQA and DUDE, the improvements demonstrate the effectiveness of layout-aware rewards in capturing cross-page dependencies. For SlideVQA, Ada-CoT helps balance direct answering and reasoning, avoiding unnecessarily long traces. Overall, these results validate that our region-aware RL framework enables stronger multi-page reasoning than prior page-level or rule-based approaches.

While the improvements over DocR1 in Table 1 are moderate, this is expected given the strong baseline and the limited sensitivity of single-page datasets to region-level grounding. In contrast, as shown in Table 2, our method achieves more consistent gains on multi-page benchmarks, where cross-page evidence aggregation is required. RegionDoc-R1 explicitly predicts question-relevant RoIs and structured reasoning traces, improving alignment between visual evidence and answers. These benefits are not fully reflected by accuracy alone, but are critical for document-level reasoning tasks.

In addition to exact-match accuracy, we also report ANLS (Average Normalized Levenshtein Similarity) scores in Appx. E. The results show consistent improvements over baseline methods, aligning with the trends observed in exact-match evaluation. We also provide single and multi-page reasoning visualizations in Appx. E.

Method	Prec.	Rec.	F1	Cov.
SFT only	58.4	54.1	56.2	60.3
GRPO	63.7	59.2	61.4	65.8
w/o R_{roi}	61.5	57.3	59.3	63.1
RA-GRPO	70.6	67.8	69.2	73.1

Table 3: Inference-time RoI grounding quality on the *SR-Doc* dev set. We report IoU-based precision, recall, F1, and coverage against verified reference RoIs.

RoI Grounding Evaluation. Beyond final answer accuracy, we further evaluate whether the predicted RoIs correctly localize answer-relevant evidence. Following the Region Alignment Reward, we compute IoU-based precision, recall, F1, and coverage between predicted RoIs and verified reference RoIs on the *SR-Doc* dev set. As shown in Table 3, RA-GRPO substantially improves grounding quality over SFT and standard GRPO. Moreover, removing R_{roi} degrades all grounding metrics, showing its importance for evidence localization.

Method	DocVQA	MP-DocVQA	Avg.
SFT only	87.7	80.7	84.2
GRPO (R_{ans} only)	91.4	84.4	87.9
w/o R_{ans}	86.9	81.0	84.0
w/o R_{roi}	90.7	83.6	87.2
w/o R_{adacot}	92.8	85.5	89.2
RA-GRPO	**95.3**	**88.3**	**91.8**

Table 4: Unified ablation study of training strategies and reward components. We report representative single-page and multi-page results.

Question Type	Direct	CoT	Avg. Tok.	Acc.
Single-evidence	68.4	31.6	246	94.7
Multi-evidence	42.1	57.9	331	91.2
Cross-page	28.6	71.4	408	88.3

Table 5: Ada-CoT routing analysis. Direct / CoT denotes the percentage of selected reasoning routes.

4.2 Ablation Study

RA-GRPO and Reward Components. To better understand the contribution of each component, we conduct a unified ablation study over different training and reward configurations. All RL variants are initialized from the same SFT checkpoint. Standard GRPO with answer reward improves over SFT-only training, confirming the benefit of RL optimization (Table 4). Removing R_{ans} leads to performance close to SFT, showing that answer correctness remains the primary optimization signal. Adding the region reward further improves both single-page and multi-page performance, as indicated by the variant w/o R_{adacot}. In contrast, w/o R_{roi} performs slightly worse than standard GRPO, suggesting that Ada-CoT without region-level grounding provides limited additional benefit. Together with Table 3, these results show that R_{roi} improves both evidence localization and task performance, while Ada-CoT serves as an auxiliary routing regularizer.

Ada-CoT Routing Analysis. We examine whether Ada-CoT adaptively controls reasoning depth. We group samples into single-evidence, multi-evidence, and cross-page questions, and report routing ratios and average reasoning length in Table 5. Ada-CoT selects direct answering more for single-evidence questions, using CoT more frequently for multi-evidence and cross-page cases. Thus, Ada-CoT adjusts reasoning depth with evidence complexity, rather than simply shortening outputs.

***SR-Doc* Dataset.** We also study the effect of our proposed SR-Doc dataset during reinforcement learning. Specifically, we replace SR-Doc with VisualCoT (Shao et al., 2024a) in the RL stage, while keeping the SFT initialization unchanged.

Training Data for RL	DocVQA	MP-DocVQA	Avg.
VisualCoT (default)	86.1	79.4	82.8
SR-Doc (ours)	**95.3**	**88.3**	**91.8**

Table 6: Ablation study on the effect of the SR-Doc dataset during RL training. Using SR-Doc leads to consistent improvements, especially in multi-page settings.

Replacing SR-Doc with VisualCoT leads to clear performance degradation, particularly on MP-DocVQA, as shown in Table 6. This is expected since VisualCoT mainly provides single-page visual reasoning with limited layout variety, lacking fine-grained RoI annotations across pages. Moreover, applying RL on general VisualCoT without document-specific region supervision may provide less reliable optimization signals for document reasoning. By contrast, SR-Doc introduces region-level supervision and multi-page coverage, which enables RA-GRPO to exploit spatial cues and align reasoning with document structure. These results highlight the importance of SR-Doc for robust layout-aware document understanding.

We further assess *SR-Doc* minimality via leave-one-out region ablation showing that annotated RoIs usually capture key evidence (Appx. F). Additional results, including analyses of hyperparameters, *SR-Doc* human audit, hallucinations/failures, and backbone generalization are in Appx. G.

5 Conclusion

We presented ***RegionDoc-R1***, a novel reinforcement learning framework for document understanding that integrates layout awareness and adaptive reasoning into MLLMs. Our approach introduces *RA-GRPO*, which grounds reasoning at the region level with explicit spatial supervision, supported by the *SR-Doc* dataset containing fine-grained multi-page annotations. In addition, the *Ada-CoT* strategy enables dynamic adjustment of reasoning depth, balancing direct answering and step-wise inference. Experiments on diverse single- and multi-page benchmarks show that RegionDoc-R1 achieves state-of-the-art performance, and demonstrates the effectiveness of combining layout-aware RL with adaptive reasoning for robust and interpretable document understanding.

Limitations

Although RegionDoc-R1 shows strong performance on document reasoning benchmarks, several limitations remain. The *SR-Doc* dataset relies on LLM-assisted annotation, which may introduce minor localization noise or hallucination-like RoI errors despite the verification step. In addition, the advantages of region-level grounding are less pronounced for tasks that rely primarily on global textual context rather than localized visual evidence. Furthermore, our current evaluation focuses on single- and multi-page documents. Extending our high-resolution, region-aware reasoning to significantly longer documents remains a key area for exploration. In future work, we plan to improve the robustness of region annotations through stronger verification strategies, explore efficient mechanisms to process long documents, and better integrate localized grounding with global document context. Moreover, our main experiments are conducted on Qwen2.5-VL-7B, with additional backbone validation on Qwen3-VL-8B in the Appendix. Extending RA-GRPO to larger MLLM backbones is a natural next step.

Acknowledgments

This work uses publicly available datasets, codebases, and pretrained models for academic research and experimental comparison. We acknowledge and cite the original creators of these artifacts in the relevant sections, and use them in accordance with their licenses, terms of use, and intended research purposes. Our newly constructed dataset is derived from existing public research datasets rather than newly collected personal data. We follow the original access conditions of the source datasets and check the derived data to avoid introducing additional personally identifying information or offensive content. We document the dataset sources, construction pipeline, statistics, annotation format, and intended use in Section 3 and Appendix C. The data construction and verification process involves annotations and quality checks conducted internally by the authors of this paper, following the annotation and verification protocol described in Section 3 and Appendix C. No external crowdworkers, paid participants, or human-subject experiments were involved. The resulting artifacts are intended solely for academic research and evaluation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 993–1003.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Aniket Bhattacharyya, Anurag Tripathi, Ujjal Das, Archan Karmakar, Amit Pathak, and Maneesh Gupta. 2025. Information extraction from visually rich documents using LLM-based organization of documents into independent textual segments. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17241–17256.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 2898–2904.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2026. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137.

Łukasz Garncarek, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama, Michał Turski, and Filip Graliński. 2021. Lambert: Layout-aware language modeling for information extraction. In *International conference on document analysis and recognition*, pages 532–547. Springer.

Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova, and Tong Sun. 2021. Unidoc: Unified pretraining framework for document understanding. *Advances in Neural Information Processing Systems*, 34:39–50.

Zhangxuan Gu, Changhua Meng, Ke Wang, Jun Lan, Weiqiang Wang, Ming Gu, and Liqing Zhang. 2022. Xylayoutlm: Towards layout-aware multimodal networks for visually-rich document understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4583–4592.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2025. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5817–5834.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. 2026. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *Advances in Neural Information Processing Systems*, 38:162239–162262.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM international conference on multimedia*, pages 4083–4091.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Anastassia Kornilova and Vladimir Eidelman. 2019. Billsum: A corpus for automatic summarization of us legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56.

Chen-Yu Lee, Chun-Liang Li, Timothy Dozat, Vincent Perot, Guolong Su, Nan Hua, Joshua Ainslie, Renshen Wang, Yasuhisa Fujii, and Tomas Pfister. 2022. Formnet: Structural encoding beyond sequential modeling in form document information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3735–3754.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387.

Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. 2020. Docbank: A benchmark dataset for document layout analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 949–960.

Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. 2021. Structext: Structured text understanding with multi-modal transformers. In *Proceedings of the 29th ACM international conference on multimedia*, pages 1912–1920.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15630–15640.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279.

Ahmed Masry, Juan Rodriguez, Tianyu Zhang, Suyuchen Wang, Chao Wang, Aarash Feizi, Akshay Kalkunte Suresh, Abhay Puri, Xiangru Jian, Pierre-André Noël, and 1 others. 2026. Alignvlm: Bridging vision and language latent spaces for multimodal document understanding. *Advances in Neural Information Processing Systems*, 38:145684–145710.

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

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

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

Vincent Perot, Kai Kang, Florian Luisier, Guolong Su, Xiaoyu Sun, Ramya Sree Boppana, Zilong Wang, Zifeng Wang, Jiaqi Mu, Hao Zhang, Chen-Yu Lee, and Nan Hua. 2024. LMDX: Language model-based document information extraction and localization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15140–15168, Bangkok, Thailand. Association for Computational Linguistics.

Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. 2021. Going full-tilt boogie on document understanding with text-image-layout transformer. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 732–747. Springer.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 2383–2392.

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

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024a. Visual cot: Advancing multimodal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024b. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *European conference on computer vision*, pages 742–758. Springer.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Stacey Svetlichnaya. 2020. Deepform: Understand structured documents at scale. *Weights & Biases report*.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13636–13645.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13878–13888.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. 2023. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19254–19264.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834.

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, and 1 others. 2023. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19528–19540.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception

of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Junyu Xiong, Yonghui Wang, Weichao Zhao, Chenyu Liu, Bing Yin, Wengang Zhou, and Houqiang Li. 2026. Docr1: Evidence page-guided grpo for multi-page document understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 11178–11186.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2087–2098.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, and 1 others. 2021. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Yu, Yijuan Chen, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, and 1 others. 2026. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *Advances in Neural Information Processing Systems*, 38:29918–29952.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, and 1 others. 2023a. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, and 1 others. 2023b. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2841–2858.

Wenwen Yu, Zhibo Yang, Yuliang Liu, and Xiang Bai. 2025. Docthinker: Explainable multimodal large language models with rule-based reinforcement learning for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 837–847.

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

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and 1 others. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A Preliminary

Group Relative Policy Optimization (GRPO) (Shao et al., 2024b) is a critic-free reinforcement learning algorithm designed for large-scale language models. Unlike classical methods such as Proximal Policy Optimization (PPO) (Schulman et al., 2017), which rely on a value network to estimate returns, GRPO avoids the instability and computational burden of training a critic by comparing sampled responses within a group.

Formally, given a question q, the policy π_θ generates K candidate responses:

$$\begin{aligned} \mathcal{O} &= \{o^{(1)}, o^{(2)}, \ldots, o^{(K)}\}, \\ o^{(i)} &= (o_1^{(i)}, o_2^{(i)}, \ldots, o_{|o^{(i)}|}^{(i)}). \end{aligned} \tag{14}$$

Each response $o^{(i)}$ receives a scalar reward $R^{(i)} \in \mathbb{R}$ from a task-specific rule-based evaluator:

$$R^{(i)} = R(q, o^{(i)}). \tag{15}$$

Relative Advantage. Instead of directly using $R^{(i)}$, GRPO computes a group-normalized *advantage* for

stability:

$$\mu = \frac{1}{K}\sum_{i=1}^{K} R^{(i)},$$
$$\sigma = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(R^{(i)}-\mu)^2+\epsilon}, \tag{16}$$

where μ and σ are the mean and standard deviation of the group, and ϵ is a stabilizer. This ensures that the learning signal depends on relative performance inside the group, not on absolute reward scales.

Token-level Credit Assignment. Each response $o^{(i)}$ is an autoregressive sequence, so its group-level advantage A_i is broadcast to each decoding step:

$$A_{i,t} = A_i, \qquad t = 1, \ldots, |o^{(i)}|. \tag{17}$$

This means every token in $o^{(i)}$ shares the same sequence-level advantage.

Policy Update. The GRPO objective encourages high-probability assignment to better-than-average responses while regularizing against a reference policy π_{ref} (the SFT checkpoint). The loss is:

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_t^{(i)} \mid q, o_{<t}^{(i)})}{\varphi[\pi_\theta(o_t^{(i)} \mid q, o_{<t}^{(i)})]}, \tag{18}$$

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{K}\sum_{i=1}^{K}\frac{1}{|o^{(i)}|} \sum_{t=1}^{|o^{(i)}|}\Big(r_{i,t}(\theta)A_{i,t} - \beta\,\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\Big). \tag{19}$$

where $r_{i,t}(\theta)$ represents the policy ratio, $\varphi[\cdot]$ is the stop-gradient operator, β controls the KL regularization strength, and π_{ref} is the SFT-initialized reference policy that anchors policy updates.

Training Procedure. In practice, GRPO is typically used in a two-stage pipeline. First, the model is initialized by supervised fine-tuning (SFT) on annotated corpora, providing a stable reference policy π_{ref} and preventing collapse at the start of RL training. Then, reinforcement learning with GRPO is applied: for each input (q, D), the model samples multiple candidate outputs $\{o^{(i)}\}_{i=1}^{K}$, computes group-normalized advantages as in Eq. (3), and updates the policy according to Eq. (6). The KL term anchors π_θ close to π_{ref}, ensuring that exploration does not drift too far from the supervised initialization.

This training paradigm allows models to iteratively refine their reasoning ability: SFT provides general-purpose generation skills, while GRPO emphasizes selecting the most promising responses within each sampled group. Such a framework has proven effective in scaling reasoning-centric large language models, and in our work, it provides the foundation for extending reinforcement learning to region-aware, layout-sensitive document reasoning.

Summary. GRPO thus replaces critic estimation with group-wise comparison, yielding three benefits: (1) critic-free stability, (2) robustness to reward scaling, and (3) scalability for LLM training. These properties make GRPO a natural foundation for our extension, RA-GRPO, where task-specific multi-objective rewards further incorporate spatial layout grounding and adaptive reasoning strategies.

B RA-GRPO Training Algorithm

To explicitly ground reasoning in semantically relevant document regions, our RA-GRPO extends GRPO by incorporating multi-objective rewards: answer correctness, region alignment, and reasoning quality. The pseudo code in Algorithm 1 summarizes the entire training loop.

For reproducibility, we summarize the **hyperparameter settings** used in RA-GRPO training:

- **Sampling:** group size $K = 8$.
- **Answer Reward R_{ans}:** tolerance $\delta = 0.05$ for numeric answers; stability $\varepsilon = 10^{-8}$.
- **Region Reward R_{roi}:** IoU threshold $\tau = 0.5$; weights $(\alpha_1, \alpha_2, \alpha_3) = (0.5, 0.4, 0.1)$; stability $\varepsilon = 10^{-8}$.
- **Ada-CoT Reward R_{adacot}:** base CoT reward $\kappa = 0.5$; length threshold $B_0 = 6$; decay factor $\delta_{\text{len}} = 3.0$; weights $(\rho_1, \rho_2) = (0.7, 0.3)$.
- **Final Reward Aggregation:** weights $(\lambda_{\text{ans}}, \lambda_{\text{roi}}, \lambda_{\text{adacot}}) = (0.4, 0.4, 0.2)$.
- **Loss Regularization:** KL penalty $\beta = 0.1$; advantage stability $\epsilon = 10^{-8}$.

C Details of *SR-Doc* dataset

Existing datasets usually lack question-specific region-level layouts and cross-page annotations, limiting their suitability for reinforcement learning

Algorithm 1: Region-Aware Group Relative Policy Optimization (RA-GRPO)

Input: Document D, question q, ground-truth answer y, ground-truth RoIs $\mathcal{G}$
Init: Policy model π_θ, reference policy π_{ref}, group size K
for *each training step* **do**
  `// Step 1: Sample candidate responses`
  Sample K responses $\{(o^{(i)}, \hat{\mathcal{B}}^{(i)}, z^{(i)}, \tau^{(i)})\}_{i=1}^{K}$ from π_θ ;
  `// Step 2: Compute rewards for each candidate`
  for $i = 1$ **to** K **do**
    `// Answer Accuracy Reward`
    if y *is textual* **then**
      $R_{\text{ans}}^{(i)} = \mathbb{1}[\text{norm}(o_{\text{final}}^{(i)}) = \text{norm}(y)]$
    else
      $e = \frac{|o_{\text{final}}^{(i)} - y|}{\max(|y|,1)}$;
      $R_{\text{ans}}^{(i)} = \exp(-e/\delta)$
    `// Region Alignment Reward`
    Compute precision, recall, and F_1 by IoU-matching $\hat{\mathcal{B}}^{(i)}$ with $\mathcal{G}$;
    Compute coverage Cov across pages ;
    $R_{\text{roi}}^{(i)} = \alpha_1 F_1 + \alpha_2 \text{Cov} - \alpha_3 \cdot \frac{\max(|\hat{\mathcal{B}}^{(i)}| - |\mathcal{G}|, 0)}{\max(|\hat{\mathcal{B}}^{(i)}|, 1)}$
    `// Adaptive CoT Reward`
    if $z^{(i)} = \text{direct}$ *and correct* **then**
      $R_{\text{gate}}^{(i)} = 1$
    else
      $R_{\text{gate}}^{(i)} = \kappa$
    $R_{\text{len}}^{(i)} = \exp(-\frac{\max(|\tau^{(i)}| - B_0, 0)}{\delta_{\text{len}}})$;
    $R_{\text{adacot}}^{(i)} = \rho_1 R_{\text{gate}}^{(i)} + \rho_2 R_{\text{len}}^{(i)}$
    `// Final Reward`
    $R^{(i)} = \lambda_{\text{ans}} R_{\text{ans}}^{(i)} + \lambda_{\text{roi}} R_{\text{roi}}^{(i)} + \lambda_{\text{adacot}} R_{\text{adacot}}^{(i)}$
  `// Step 3: Compute group-normalized advantages`
  $\mu = \frac{1}{K}\sum_i R^{(i)}, \sigma = \sqrt{\frac{1}{K}\sum_i (R^{(i)} - \mu)^2 + \epsilon}$;
  $A_i = (R^{(i)} - \mu)/\sigma$ for each i
  `// Step 4: Policy update`
  Update π_θ by minimizing:

$$\mathcal{L}_{\text{RA-GRPO}} = -\frac{1}{K}\sum_{i=1}^{K}\frac{1}{|o^{(i)}|}\sum_{t=1}^{|o^{(i)}|}\left(\frac{\pi_\theta(o_t^{(i)} \mid q, D, o_{<t}^{(i)})}{\varphi[\pi_\theta(o_t^{(i)} \mid q, D, o_{<t}^{(i)})]} \cdot A_i - \beta\, \mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\right)$$

in document understanding. To address this gap, we construct *SR-Doc*, a hybrid corpus designed for region-aware reinforcement learning, by aggregating a broad range of existing document understanding datasets. For single-page documents, we use ten commonly used benchmarks, including DocVQA (Mathew et al., 2021), InfoGraphicVQA (Mathew et al., 2022), ChartQA (Masry et al., 2022), DeepForm (Svetlichnaya, 2020), FigureQA (Kahou et al., 2017), OCRVQA (Mishra et al., 2019), TabFact (Chen et al., 2020), TextCaps (Sidorov et al., 2020), TextVQA (Singh et al., 2019), and VisualMRC (Tanaka et al., 2021). For multi-page documents, we curate three datasets: ArxivQA (Li et al., 2024), DUDE (Van Landeghem et al., 2023), and MP-DocVQA (Tito et al., 2023).

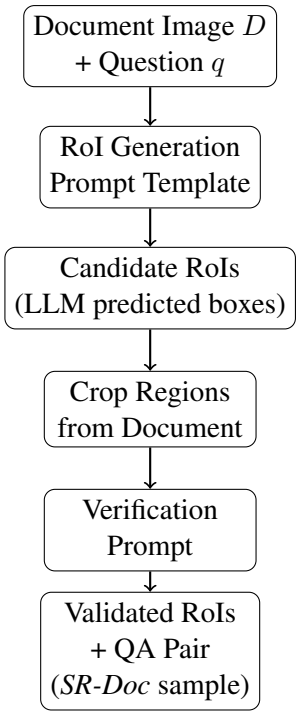


Figure 3: Pipeline for constructing the *SR-Doc* dataset. Given a document and a question, a multimodal LLM first predicts question-relevant regions of interest (RoIs). The predicted regions are then cropped and used in a verification step to ensure that the answer can be derived solely from the selected regions. Only QA pairs whose answers can be reproduced from the cropped regions are retained as *SR-Doc* annotations.

To obtain high-quality region-level supervision, we adopt a two-step annotation strategy based on Qwen2-VL-72B-Instruct (Wang et al., 2024). The goal is to obtain question-specific Regions of Interest (RoIs) that contain sufficient visual evidence for answering the question. An overview of the *SR-Doc* annotation pipeline is illustrated in Figure 3.

Stage 1: RoI Identification

Given a document image D and a question-answer pair (q, y) from an existing dataset, we first prompt the model to identify the regions most relevant for answering the question. The prompt encourages the model to perform step-by-step reasoning and output bounding boxes corresponding to the evidence regions. The prompt template is shown below:

> **Instruction:** You are given a document image and a question about the document. Identify the regions in the document that provide evidence for answering the question.
>
> **Input:**
>
> - Document image
> - Question: `{question}`
> - Ground-truth answer: `{answer}`
>
> **Output format:**
>
> ```
> <think>
> Explain briefly how you locate the
> evidence in the document.
> </think>
>
> <answer>
> {
>  "bboxes": [[x1, y1, x2, y2], ...],
>  "final answer": "{answer}"
> }
> </answer>
> ```

The model is instructed to first reason about where the relevant information appears in the document and then output the bounding boxes corresponding to the supporting regions. This step produces candidate RoIs that are semantically aligned with the question.

Stage 2: Sufficiency Verification

To ensure that the predicted RoIs contain sufficient evidence for answering the question, we perform a second verification step. Specifically, the predicted regions are cropped from the original document and fed back to the model along with the question.

The model is asked to answer the question using only the cropped regions. If the predicted answer matches the ground truth, the RoIs are considered valid and retained. The verification prompt is defined as:

> **Instruction:** You are given cropped regions from a document. Use only the provided regions to answer the question.
>
> **Input:**
>
> - Cropped document regions
> - Question: `{question}`
>
> **Output format:**
>
> ```
> <think>
> Explain how the answer can be
> derived from the given regions.
> </think>
>
> <answer>
> ```

```
{predicted answer}
</answer>
```

If the predicted answer matches the ground-truth answer after normalization, the corresponding QA–RoI pair is retained. Otherwise, the annotation is discarded. For multi-page samples, the model is allowed to select and validate RoIs across different pages, yielding explicit cross-page evidence supervision that is not provided in the original datasets.

Overall, *SR-Doc* contains approximately **130k documents** with around **450k QA pairs**, aggregated from both single-page and multi-page sources. On average, each document contains **3.5 pages**, with each QA pair annotated by **4.1 RoIs**, reflecting the localized evidence needed for different questions. We split the dataset into 120k training and 10k dev samples to validate training stability.

In conclusion, *SR-Doc* provides two main characteristics: **(1) Region-level RoIs.** Instead of coarse page-level grounding, each question is paired with fine-grained RoIs directly linked to answer semantics, enabling precise layout-aware supervision. **(2) Cross-page grounding.** Many question-relevant RoIs span multiple pages, converting existing multi-page QA pairs into RoI-grounded supervision for training.

D More Training Details

Our RegionDoc-R1 is built on top of Qwen2.5-VL-7B and trained on 8 NVIDIA H100 GPUs with mixed-precision (FP16) optimization. All experiments are implemented in PyTorch. We adopt a two-stage training strategy. In the SFT stage, we train on Visual CoT, which contains 438K QA pairs, for 3 epochs with a batch size of 128. Images are resized to a maximum resolution of 896×896 while preserving aspect ratio, and textual inputs are truncated to 512 tokens. We use AdamW with a learning rate of 2e-5, a cosine decay schedule, and weight decay of 0.01. In the RL stage, we fine-tune on *SR-Doc*, which contains around 450K QA pairs, for 1.5 epochs with a batch size of 64. The maximum input resolution is increased to 1024×1024 to better capture fine-grained RoIs. We use AdamW with a peak learning rate of 1e-5 and apply gradient clipping of 1.0 to stabilize policy updates. Training takes approximately 6 days for the SFT stage and 4 days for the RL stage.

Compared with standard GRPO, RA-GRPO introduces additional overhead from RoI matching and multi-objective reward computation, but it does not require an additional critic model. Thus, the memory cost remains comparable to standard GRPO, while the main extra cost comes from computing IoU-based region rewards. For comparison, VisCoT is trained on Visual CoT, DocThinker reports using 8K document RL data, while the exact training scale of some baselines is not clearly reported in their original papers.

Model	DocVQA	InfoVQA	MP-DocVQA
Qwen2.5-VL	94.6	81.5	87.9
DocR1	95.5	83.1	88.2
RegionDoc-R1	**95.8**	**83.9**	**89.1**

Table 7: Evaluation using ANLS. The results follow similar trends to the exact-match accuracy reported.

E More Results

ANLS Evaluation. Following common practice in document question answering, we additionally evaluate our model using ANLS (Average Normalized Levenshtein Similarity), which accounts for minor string variations caused by OCR errors or formatting inconsistencies.

As shown in Table 7, the ANLS scores are consistent with the trends observed in the exact-match evaluation reported in the main paper. RegionDoc-R1 achieves the best performance across all datasets, confirming that the improvements are not limited to strict exact-match metrics but also hold under relaxed string similarity evaluation. These results further demonstrate that the proposed region-aware reinforcement learning framework improves document reasoning robustness while maintaining strong answer accuracy.

Qualitative Examples. Figure 4 shows our method consistently identifies accurate RoIs and produces coherent reasoning traces across diverse document types. These visual examples highlight how region-level grounding and adaptive CoT jointly enable precise and interpretable document understanding. Furthermore, we provide cross-page examples in Figure 5 and Figure 6 to illustrate its region-level grounding and reasoning behavior on diverse multi-page documents. As shown in these figures, the model consistently identifies question-relevant RoIs and generates coherent reasoning traces that link visual evidence to the final answer. In most cases, the predicted regions span multiple pages or distributed layout elements, demonstrating the ability of RegionDoc-R1 to aggregate evidence from

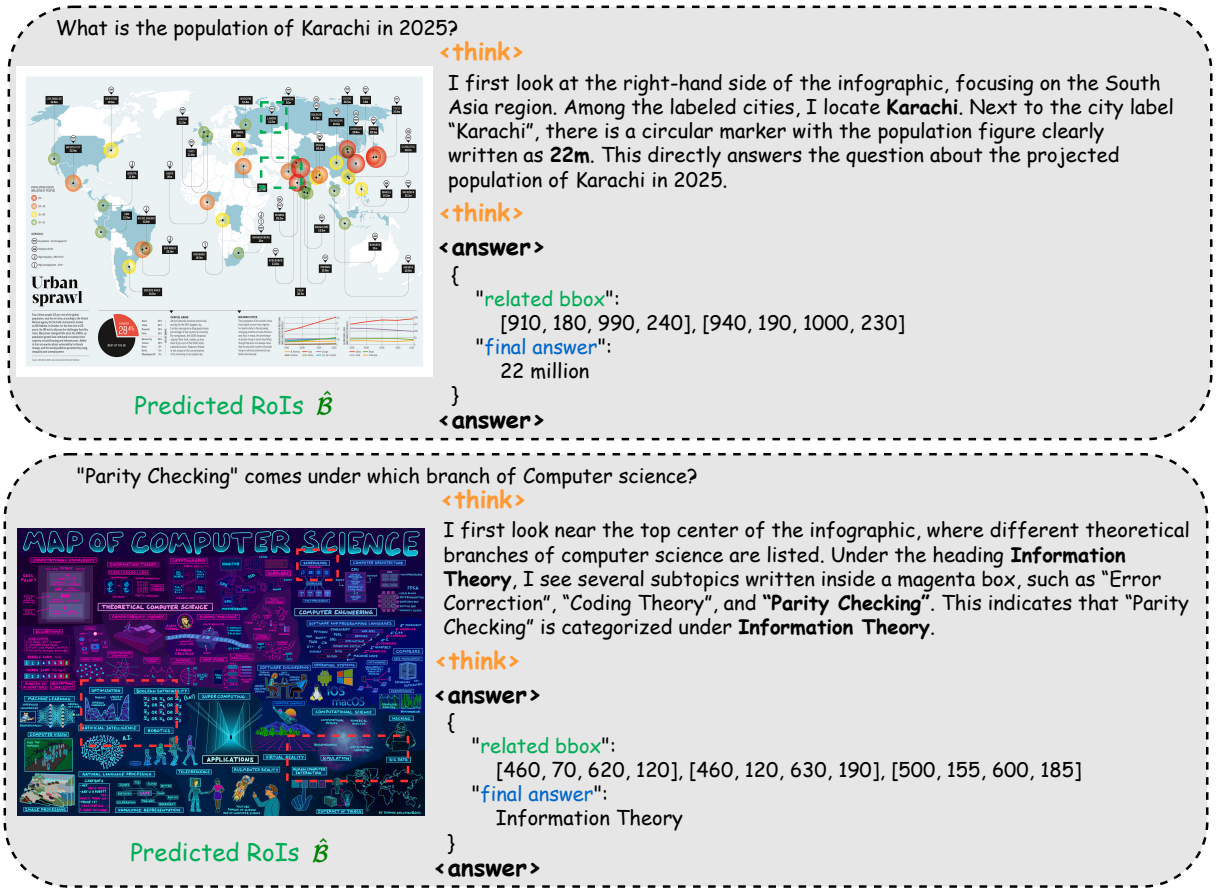


Figure 4: **Inference cases on single-page documents from InfoVQA.** The thinking process significantly improves the reasoning ability and explainability. Please zoom in to see the details.

Setting	DocVQA	MP-DocVQA
All RoIs	95.3	88.3
Remove one RoI	84.6	76.1
Remove two RoIs	71.4	63.8

Table 8: Leave-one-out region ablation on *SR-Doc* annotations. "All RoIs" denotes answering with all verified reference RoIs as cropped evidence. Removing individual RoIs significantly reduces answer accuracy, suggesting that the annotated regions capture key evidence required for answering the question.

spatially separated document regions.

F Minimality Analysis of *SR-Doc*

To further analyze whether the annotated RoIs correspond to minimal evidence regions, we conduct a leave-one-out region ablation study on *SR-Doc*.

For each QA sample with a set of annotated RoIs $\mathcal{G} = \{g_1, g_2, \ldots, g_n\}$, we first evaluate the *All RoIs* setting, where all verified reference RoIs are provided as cropped evidence. We then construct ablated inputs by removing one or two RoIs and evaluate whether the correct answer can still be obtained.

As shown in Table 8, removing a single region already leads to a noticeable drop in accuracy, particularly for multi-page documents. This indicates that the annotated RoIs are closely tied to the evidence required for answering the question. When multiple regions are removed, performance degrades further, suggesting that many questions rely on complementary evidence distributed across different regions or pages. Overall, these results provide empirical support that the RoIs annotated in *SR-Doc* capture key evidence for document reasoning and are not redundant layout elements.

G Additional Analysis

G.1 Hyperparameters

We further analyze the sensitivity of RegionDoc-R1 to several key hyperparameters in RA-GRPO training, including the region reward weight λ_{roi}, the reasoning-length threshold B_0 in Ada-CoT, and the GRPO group size K. Results are given in Table 9. Unless otherwise specified, all experiments follow the same training and evaluation settings as

Hyperparameter	Value	DocVQA	MP-DocVQA
λ_{roi}	0.1	94.7	87.2
	0.3	95.1	87.8
	0.4	**95.3**	**88.3**
	0.5	95.2	88.1
	0.7	95.0	87.9
B_0	4	94.9	87.6
	6	**95.3**	**88.3**
	8	95.2	88.1
	10	95.0	87.9
K	4	94.6	87.4
	8	**95.3**	**88.3**
	16	95.4	88.2

Table 9: Sensitivity analysis of key hyperparameters in RA-GRPO. We vary one hyperparameter at a time while keeping the others fixed. The results show that RegionDoc-R1 is relatively stable across a reasonable range of values, with the best overall performance achieved by $\lambda_{\text{roi}} = 0.4$, $B_0 = 6$, and $K = 8$.

the main paper, while varying only one hyperparameter at a time.

Effect of λ_{roi}. We vary the weight of the region-aware reward while keeping the remaining reward weights fixed. Introducing region-level supervision consistently improves performance over smaller values, especially on MP-DocVQA, where cross-page grounding is more critical. However, assigning an excessively large weight to R_{roi} slightly hurts answer accuracy, likely because the optimization becomes overly constrained by spatial alignment. We therefore adopt $\lambda_{\text{roi}} = 0.4$ in the final model.

Effect of B_0. The Ada-CoT reward softly penalizes reasoning traces longer than the threshold B_0. As shown in Table 9, moderate reasoning length yields the best performance. When B_0 is too small, the model may be discouraged from performing sufficient multi-step reasoning; when B_0 is too large, reasoning traces become unnecessarily long and less efficient. We therefore set $B_0 = 6$.

Effect of group size K. We also study the number of sampled candidates used in GRPO training. Increasing K improves reward estimation stability, but also increases training cost. In our experiments, moving from $K = 4$ to $K = 8$ yields a clear gain, while further increasing to $K = 16$ brings only marginal improvement. We thus choose $K = 8$ as a good trade-off between performance and efficiency.

Overall, RegionDoc-R1 is relatively robust to moderate hyperparameter variations. The results indicate that the proposed RA-GRPO framework does not rely on delicate tuning, and achieves stable performance across a reasonable range of settings.

Split	Valid	Partially Valid	Invalid
Single-page	88.6	8.4	3.0
Multi-page	84.2	10.8	5.0
Overall	86.9	9.3	3.8

Table 10: Human audit of *SR-Doc* RoI annotation quality. We randomly sample 500 QA–RoI pairs and manually inspect whether the annotated RoIs contain sufficient evidence to answer the question.

G.2 Human Audit of SR-Doc Annotations

Although *SR-Doc* uses LLM-assisted RoI annotations, the generated regions are not directly used as training supervision. As described in Appendix. C, each candidate QA–RoI pair is first filtered by crop-based sufficiency verification, where the cropped RoIs must reproduce the ground-truth answer. Here, the reference RoIs used for grounding evaluation are the verified *SR-Doc* annotations, which are generated by Qwen2-VL-72B-Instruct and filtered by crop-based sufficiency verification. To further assess the reliability of these verified reference RoIs, we conduct a manual audit on randomly sampled QA–RoI pairs from the *SR-Doc* dev set.

Specifically, we randomly sample 500 QA–RoI pairs from the *SR-Doc* dev set, including both single-page and multi-page documents. Each sample is manually inspected and categorized into one of three groups: *Valid*, where the annotated RoIs contain sufficient evidence to answer the question; *Partially Valid*, where the RoIs are relevant but miss some supporting context or include loose boundaries; and *Invalid*, where the RoIs are irrelevant or insufficient for answering the question.

As shown in Table 10, most annotations are judged as valid, indicating that the proposed annotation pipeline produces reliable region-level supervision. Multi-page samples show a slightly higher partially valid or invalid rate, mainly because evidence may be distributed across distant pages or multiple layout elements. The invalid cases are mostly caused by missing auxiliary evidence, overly broad regions, or occasional teacher-model localization errors. Overall, the human audit, together with the crop-based sufficiency verification and leave-one-out ablation in Sec. F, supports the reliability of *SR-Doc* annotations.

Method	Acc.	Unsupported Ans. ↓	Grounded Acc. ↑
SFT only	84.2	18.6	68.4
GRPO	87.9	15.2	74.6
w/o R_{roi}	87.2	16.1	73.2
RA-GRPO	91.8	8.7	84.6

Table 11: Hallucination and grounded accuracy analysis on 500 randomly sampled prediction results. Unsupported answers refer to predictions whose final answers are not sufficiently supported by the predicted RoIs.

Failure Type	Ratio
Incorrect / missing RoI	34.7
Incomplete cross-page evidence	24.3
Correct RoI but wrong reasoning	18.0
Incorrect direct routing	12.7
Formatting / extraction error	10.3

Table 12: Failure case and CoT error propagation analysis on 300 randomly sampled incorrect predictions.

G.3 Hallucination and Grounded Accuracy Analysis

We further analyze whether region-aware reinforcement learning reduces hallucinated or unsupported answers. Since automatic LLM-based judging may introduce circular evaluation, we conduct a manual audit on randomly sampled prediction results. Specifically, we sample 500 question-answer instances from representative document reasoning benchmarks, including both single-page and multi-page samples. For each instance, annotators inspect the document, the predicted RoIs, and the final answer, and determine whether the predicted RoIs contain sufficient evidence to support the answer. Two annotators independently inspect each sampled prediction, and disagreements are resolved through discussion.

We report two metrics: *Unsupported Answer Rate*, which measures the proportion of predictions whose final answers are not supported by the predicted RoIs, and *Grounded Accuracy*, which measures the proportion of predictions that are both correct and supported by the predicted RoIs.

As shown in Table 11, RA-GRPO achieves the lowest unsupported answer rate and the highest grounded accuracy. Compared with standard GRPO, RA-GRPO reduces unsupported answers by encouraging the model to localize evidence before producing the final answer. The variant without R_{roi} performs worse than full RA-GRPO, indicating that answer-level RL alone cannot fully prevent the model from relying on shortcut or unsupported predictions. These results suggest that region-aware rewards help align final answers with localized document evidence and reduce hallucination-like behaviors.

G.4 Failure Case and CoT Error Propagation Analysis

To better understand the limitations of RegionDoc-R1, we further analyze failure cases where the final answer is incorrect. We randomly sample 300 incorrect predictions from representative single-page and multi-page benchmarks and manually categorize the main source of error by inspecting the predicted RoIs, intermediate reasoning trace, and final answer.

We consider five common failure types: (1) *Incorrect or missing RoI*, where the model fails to localize the key evidence region; (2) *Incomplete cross-page evidence*, where the model identifies part of the evidence but misses supporting regions on other pages; (3) *Correct RoI but wrong reasoning*, where the model localizes the relevant evidence but performs incorrect comparison, counting, or semantic inference; (4) *Incorrect direct routing*, where the model chooses direct answering for questions that require multi-step reasoning; and (5) *Formatting or extraction error*, where the model identifies the correct evidence but outputs the answer in an incompatible format.

As shown in Table 12, the most common failures are caused by incorrect or incomplete evidence localization, especially when relevant information is distributed across multiple pages. This suggests that grounding remains a key bottleneck for document reasoning. We also observe cases where the predicted RoIs are correct but the reasoning trace is flawed, indicating that accurate localization does not always guarantee correct reasoning. Incorrect direct routing accounts for a smaller but non-negligible portion of errors, suggesting that Ada-CoT can still misjudge question complexity. Overall, these results show that errors may propagate from faulty localization to reasoning and final prediction, while also highlighting the need for stronger reasoning verification in future work.

G.5 Backbone Generalization Analysis

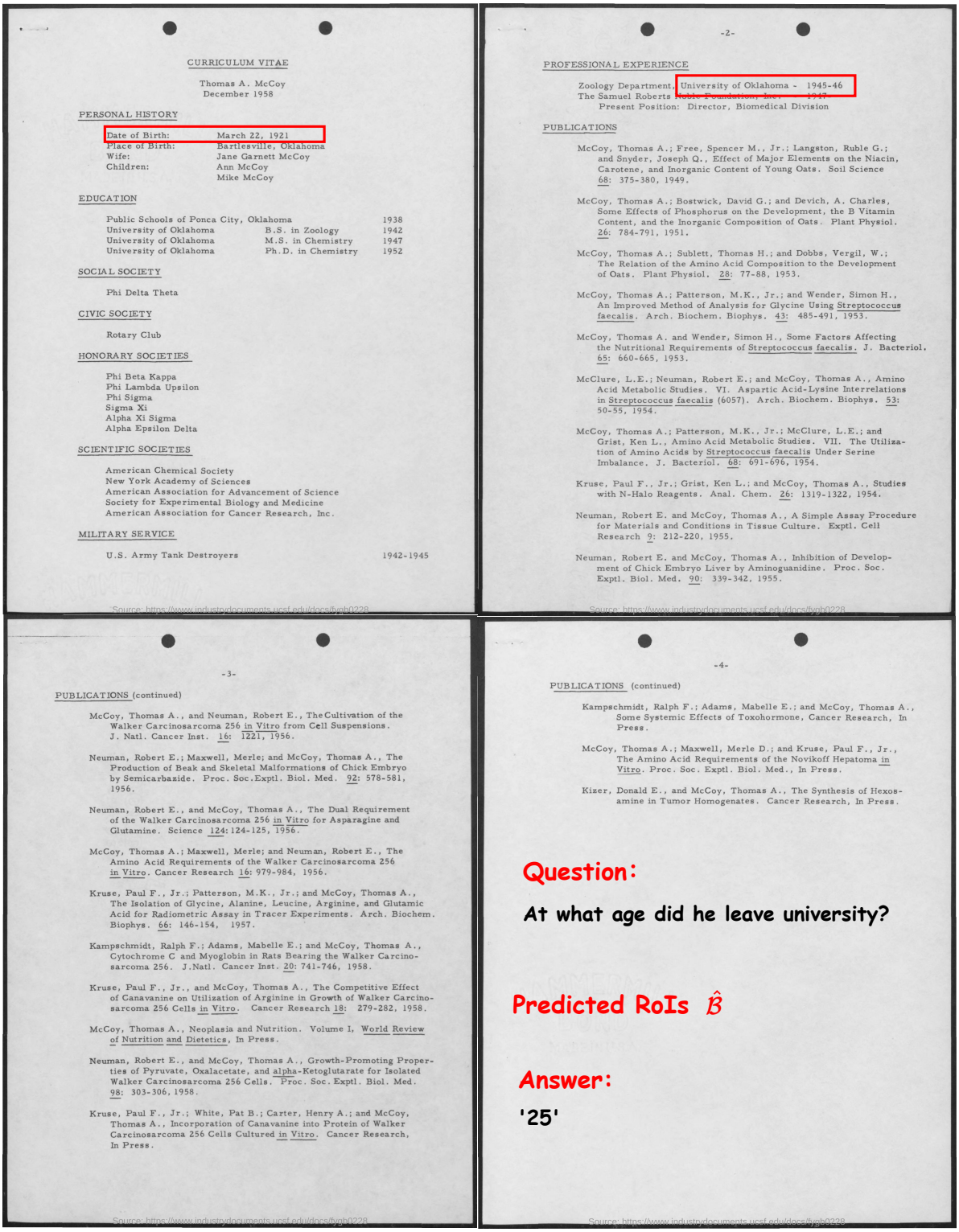
To examine whether RA-GRPO generalizes beyond Qwen2.5-VL (Wang et al., 2024), we further conduct experiments on Qwen3-VL-8B (Bai et al., 2025). We evaluate representative single-

Backbone	Method	DocVQA	MP-DocVQA
Qwen2.5-VL-7B	GRPO	91.4	84.4
Qwen2.5-VL-7B	RA-GRPO	95.3	88.3
Qwen3-VL-8B	GRPO	92.6	85.1
Qwen3-VL-8B	RA-GRPO	96.1	89.0

Table 13: Backbone generalization analysis on representative single-page and multi-page benchmarks. RA-GRPO consistently improves over standard GRPO on both Qwen2.5-VL-7B and Qwen3-VL-8B.

page and multi-page document reasoning benchmarks, including DocVQA and MP-DocVQA. For each backbone, we compare standard GRPO with our RA-GRPO under the same training and evaluation protocol.

As shown in Table 13, RA-GRPO consistently improves over standard GRPO on both Qwen2.5-VL-7B and Qwen3-VL-8B. These results suggest that the proposed region-aware reward design is not tied to a specific backbone, but can generalize to stronger MLLMs that support document understanding and structured RoI prediction.



CURRICULUM VITAE

Thomas A. McCoy
December 1958

PERSONAL HISTORY

Date of Birth:	March 22, 1921
Place of Birth:	Bartlesville, Oklahoma
Wife:	Jane Garnett McCoy
Children:	Ann McCoy
	Mike McCoy

EDUCATION

Public Schools of Ponca City, Oklahoma		1938
University of Oklahoma	B.S. in Zoology	1942
University of Oklahoma	M.S. in Chemistry	1947
University of Oklahoma	Ph.D. in Chemistry	1952

SOCIAL SOCIETY

Phi Delta Theta

CIVIC SOCIETY

Rotary Club

HONORARY SOCIETIES

Phi Beta Kappa
Phi Lambda Upsilon
Phi Sigma
Sigma Xi
Alpha Xi Sigma
Alpha Epsilon Delta

SCIENTIFIC SOCIETIES

American Chemical Society
New York Academy of Sciences
American Association for Advancement of Science
Society for Experimental Biology and Medicine
American Association for Cancer Research, Inc.

MILITARY SERVICE

U.S. Army Tank Destroyers 1942-1945

Source: https://www.industrydocuments.ucsf.edu/docs/fygb0228

-2-

PROFESSIONAL EXPERIENCE

Zoology Department, University of Oklahoma - 1945-46
The Samuel Roberts Noble Foundation, Inc. - 1947-
Present Position: Director, Biomedical Division

PUBLICATIONS

McCoy, Thomas A.; Free, Spencer M., Jr.; Langston, Ruble G.; and Snyder, Joseph Q., Effect of Major Elements on the Niacin, Carotene, and Inorganic Content of Young Oats. Soil Science 68: 375-380, 1949.

McCoy, Thomas A.; Bostwick, David G.; and Devich, A. Charles, Some Effects of Phosphorus on the Development, the B Vitamin Content, and the Inorganic Composition of Oats. Plant Physiol. 26: 784-791, 1951.

McCoy, Thomas A.; Sublett, Thomas H.; and Dobbs, Vergil, W.; The Relation of the Amino Acid Composition to the Development of Oats. Plant Physiol. 28: 77-88, 1953.

McCoy, Thomas A.; Patterson, M.K., Jr.; and Wender, Simon H., An Improved Method of Analysis for Glycine Using Streptococcus faecalis. Arch. Biochem. Biophys. 43: 485-491, 1953.

McCoy, Thomas A. and Wender, Simon H., Some Factors Affecting the Nutritional Requirements of Streptococcus faecalis. J. Bacteriol. 65: 660-665, 1953.

McClure, L.E.; Neuman, Robert E.; and McCoy, Thomas A., Amino Acid Metabolic Studies. VI. Aspartic Acid-Lysine Interrelations in Streptococcus faecalis (6057). Arch. Biochem. Biophys. 53: 50-55, 1954.

McCoy, Thomas A.; Patterson, M.K., Jr.; McClure, L.E.; and Grist, Ken L., Amino Acid Metabolic Studies. VII. The Utilization of Amino Acids by Streptococcus faecalis Under Serine Imbalance. J. Bacteriol. 68: 691-696, 1954.

Kruse, Paul F., Jr.; Grist, Ken L.; and McCoy, Thomas A., Studies with N-Halo Reagents. Anal. Chem. 26: 1319-1322, 1954.

Neuman, Robert E. and McCoy, Thomas A., A Simple Assay Procedure for Materials and Conditions in Tissue Culture. Exptl. Cell Research 9: 212-220, 1955.

Neuman, Robert E. and McCoy, Thomas A., Inhibition of Development of Chick Embryo Liver by Aminoguanidine. Proc. Soc. Exptl. Biol. Med. 90: 339-342, 1955.

Source: https://www.industrydocuments.ucsf.edu/docs/fygb0228

-3-

PUBLICATIONS (continued)

McCoy, Thomas A., and Neuman, Robert E., The Cultivation of the Walker Carcinosarcoma 256 in Vitro from Cell Suspensions. J. Natl. Cancer Inst. 16: 1221, 1956.

Neuman, Robert E.; Maxwell, Merle; and McCoy, Thomas A., The Production of Beak and Skeletal Malformations of Chick Embryo by Semicarbazide. Proc. Soc.Exptl. Biol. Med. 92: 578-581, 1956.

Neuman, Robert E., and McCoy, Thomas A., The Dual Requirement of the Walker Carcinosarcoma 256 in Vitro for Asparagine and Glutamine. Science 124: 124-125, 1956.

McCoy, Thomas A.; Maxwell, Merle; and Neuman, Robert E., The Amino Acid Requirements of the Walker Carcinosarcoma 256 in Vitro. Cancer Research 16: 979-984, 1956.

Kruse, Paul F., Jr.; Patterson, M.K., Jr.; and McCoy, Thomas A., The Isolation of Glycine, Alanine, Leucine, Arginine, and Glutamic Acid for Radiometric Assay in Tracer Experiments. Arch. Biochem. Biophys. 66: 146-154, 1957.

Kampschmidt, Ralph F.; Adams, Mabelle E.; and McCoy, Thomas A., Cytochrome C and Myoglobin in Rats Bearing the Walker Carcinosarcoma 256. J.Natl. Cancer Inst. 20: 741-746, 1958.

Kruse, Paul F., Jr., and McCoy, Thomas A., The Competitive Effect of Canavanine on Utilization of Arginine in Growth of Walker Carcinosarcoma 256 Cells in Vitro. Cancer Research 18: 279-282, 1958.

McCoy, Thomas A., Neoplasia and Nutrition. Volume I, World Review of Nutrition and Dietetics, In Press.

Neuman, Robert E., and McCoy, Thomas A., Growth-Promoting Properties of Pyruvate, Oxalacetate, and alpha-Ketoglutarate for Isolated Walker Carcinosarcoma 256 Cells. Proc. Soc. Exptl. Biol. Med. 98: 303-306, 1958.

Kruse, Paul F., Jr.; White, Pat B.; Carter, Henry A.; and McCoy, Thomas A., Incorporation of Canavanine into Protein of Walker Carcinosarcoma 256 Cells Cultured in Vitro. Cancer Research, In Press.

Source: https://www.industrydocuments.ucsf.edu/docs/fygb0228

-4-

PUBLICATIONS (continued)

Kampschmidt, Ralph F.; Adams, Mabelle E.; and McCoy, Thomas A., Some Systemic Effects of Toxohormone, Cancer Research, In Press.

McCoy, Thomas A.; Maxwell, Merle D.; and Kruse, Paul F., Jr., The Amino Acid Requirements of the Novikoff Hepatoma in Vitro. Proc. Soc. Exptl. Biol. Med., In Press.

Kizer, Donald E., and McCoy, Thomas A., The Synthesis of Hexosamine in Tumor Homogenates. Cancer Research, In Press.

Question:

At what age did he leave university?

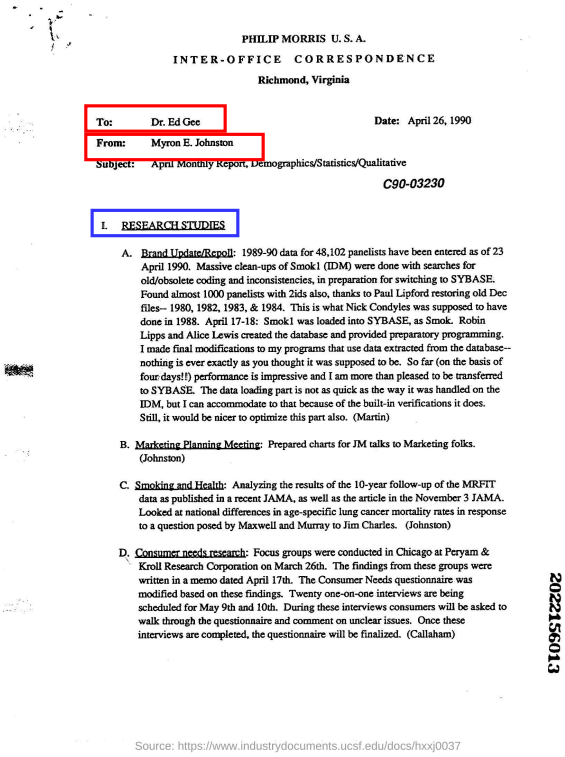
Predicted RoIs $\hat{\mathcal{B}}$

Answer:

'25'

Source: https://www.industrydocuments.ucsf.edu/docs/fygb0228

Figure 5: **Inference cases on multi-page documents from MP-DocVQA.** RegionDoc-R1 understands the question, identifies the relevant regions, and correctly captures the semantic relationships among these regions based on the question. Please zoom in to see the details.



PHILIP MORRIS U. S. A.

INTER-OFFICE CORRESPONDENCE

Richmond, Virginia

To: Dr. Ed Gee

Date: April 26, 1990

From: Myron E. Johnston

Subject: April Monthly Report, Demographics/Statistics/Qualitative

C90-03230

I. RESEARCH STUDIES

A. Brand Update/Repoll: 1989-90 data for 48,102 panelists have been entered as of 23 April 1990. Massive clean-ups of Smok1 (IDM) were done with searches for old/obsolete coding and inconsistencies, in preparation for switching to SYBASE. Found almost 1000 panelists with 2ids also, thanks to Paul Lipford restoring old Dec files-- 1980, 1982, 1983, & 1984. This is what Nick Condyles was supposed to have done in 1988. April 17-18: Smok1 was loaded into SYBASE, as Smok. Robin Lipps and Alice Lewis created the database and provided preparatory programming. I made final modifications to my programs that use data extracted from the database-- nothing is ever exactly as you thought it was supposed to be. So far (on the basis of four days!!) performance is impressive and I am more than pleased to be transferred to SYBASE. The data loading part is not as quick as the way it was handled on the IDM, but I can accommodate to that because of the built-in verifications it does. Still, it would be nicer to optimize this part also. (Martin)

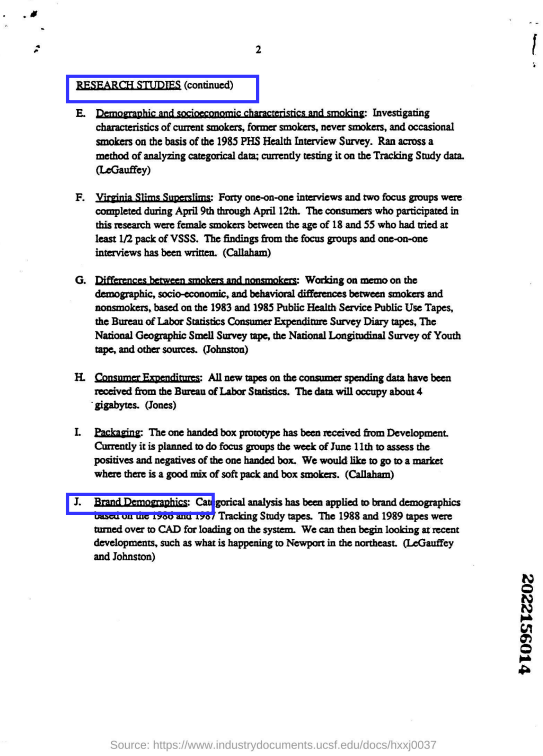
B. Marketing Planning Meeting: Prepared charts for JM talks to Marketing folks. (Johnston)

C. Smoking and Health: Analyzing the results of the 10-year follow-up of the MRFIT data as published in a recent JAMA, as well as the article in the November 3 JAMA. Looked at national differences in age-specific lung cancer mortality rates in response to a question posed by Maxwell and Murray to Jim Charles. (Johnston)

D. Consumer needs research: Focus groups were conducted in Chicago at Peryam & Kroll Research Corporation on March 26th. The findings from these groups were written in a memo dated April 17th. The Consumer Needs questionnaire was modified based on these findings. Twenty one-on-one interviews are being scheduled for May 9th and 10th. During these interviews consumers will be asked to walk through the questionnaire and comment on unclear issues. Once these interviews are completed, the questionnaire will be finalized. (Callaham)

2022156013

Source: https://www.industrydocuments.ucsf.edu/docs/hxxj0037



2

RESEARCH STUDIES (continued)

E. Demographic and socioeconomic characteristics and smoking: Investigating characteristics of current smokers, former smokers, never smokers, and occasional smokers on the basis of the 1985 PHS Health Interview Survey. Ran across a method of analyzing categorical data; currently testing it on the Tracking Study data. (LeGauffey)

F. Virginia Slims Superslims: Forty one-on-one interviews and two focus groups were completed during April 9th through April 12th. The consumers who participated in this research were female smokers between the age of 18 and 55 who had tried at least 1/2 pack of VSSS. The findings from the focus groups and one-on-one interviews has been written. (Callaham)

G. Differences between smokers and nonsmokers: Working on memo on the demographic, socio-economic, and behavioral differences between smokers and nonsmokers, based on the 1983 and 1985 Public Health Service Public Use Tapes, the Bureau of Labor Statistics Consumer Expenditure Survey Diary tapes, The National Geographic Smell Survey tape, the National Longitudinal Survey of Youth tape, and other sources. (Johnston)

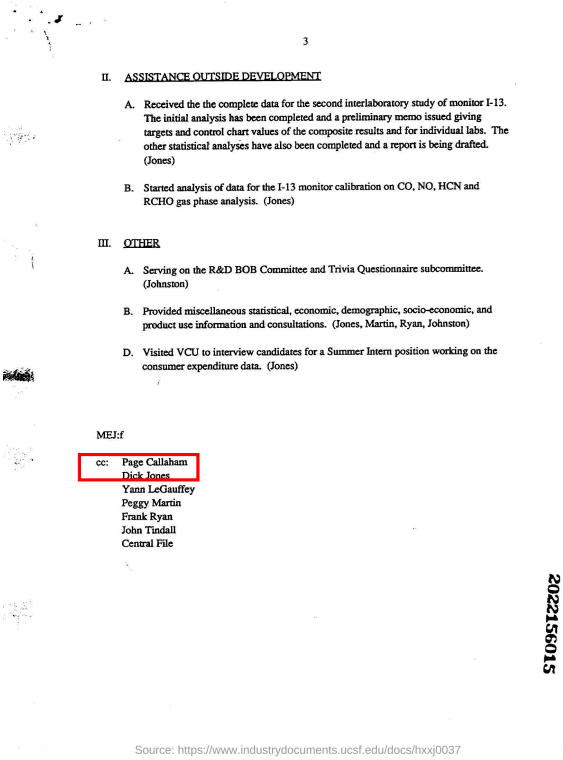
H. Consumer Expenditures: All new tapes on the consumer spending data have been received from the Bureau of Labor Statistics. The data will occupy about 4 gigabytes. (Jones)

I. Packaging: The one handed box prototype has been received from Development. Currently it is planned to do focus groups the week of June 11th to assess the positives and negatives of the one handed box. We would like to go to a market where there is a good mix of soft pack and box smokers. (Callaham)

J. Brand Demographics: Categorical analysis has been applied to brand demographics based on the 1988 and 1987 Tracking Study tapes. The 1988 and 1989 tapes were turned over to CAD for loading on the system. We can then begin looking at recent developments, such as what is happening to Newport in the northeast. (LeGauffey and Johnston)

2022156014

Source: https://www.industrydocuments.ucsf.edu/docs/hxxj0037



3

II. ASSISTANCE OUTSIDE DEVELOPMENT

A. Received the the complete data for the second interlaboratory study of monitor I-13. The initial analysis has been completed and a preliminary memo issued giving targets and control chart values of the composite results and for individual labs. The other statistical analyses have also been completed and a report is being drafted. (Jones)

B. Started analysis of data for the I-13 monitor calibration on CO, NO, HCN and RCHO gas phase analysis. (Jones)

III. OTHER

A. Serving on the R&D BOB Committee and Trivia Questionnaire subcommittee. (Johnston)

B. Provided miscellaneous statistical, economic, demographic, socio-economic, and product use information and consultations. (Jones, Martin, Ryan, Johnston)

D. Visited VCU to interview candidates for a Summer Intern position working on the consumer expenditure data. (Jones)

MEJ:f

cc: Page Callaham
Dick Jones
Yann LeGauffey
Peggy Martin
Frank Ryan
John Tindall
Central File

2022156015

Source: https://www.industrydocuments.ucsf.edu/docs/hxxj0037

Question:

Who are the recipient, the sender, and the first person in CC?

Answer:

'Dr. Ed Gee', 'Myron E. Johnston', 'Page Callaham'

Question:

What is the topic of the last research study in this report?

Answer:

'Brand Demographics'

Figure 6: **Inference cases on multi-page documents from MP-DocVQA.** QA pairs with different colors correspond to predicted ROIs of the same color. Please zoom in to see the details.