

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 INFINITY-PARSER: LAYOUT-AWARE REINFORCEMENT LEARNING FOR SCANNED DOCUMENT PARSING

Anonymous authors

Paper under double-blind review

## ABSTRACT

Document parsing from scanned images into structured formats remains a significant challenge due to its complexly intertwined elements such as text paragraphs, figures, formulas, and tables. Existing supervised fine-tuning methods often struggle to generalize across diverse document types, leading to poor performance, particularly on out-of-distribution data. This issue is further exacerbated by the limited availability of high-quality training data for layout-aware parsing tasks. To address these challenges, we introduce LayoutRL, a reinforcement learning framework that optimizes layout understanding through composite rewards integrating normalized edit distance, paragraph count accuracy, and reading order preservation. To support this training, we construct the Infinity-Doc-400K dataset, which we use to train Infinity-Parser, a vision-language model demonstrating robust generalization across various domains. Extensive evaluations on benchmarks including OmniDocBench, olmOCR-Bench, PubTabNet, and FinTabNet show that Infinity-Parser consistently achieves state-of-the-art performance across a broad range of document types, languages, and structural complexities, substantially outperforming both specialized document parsing systems and general-purpose vision-language models. We will release our code, dataset, and model to facilitate reproducible research in document parsing.

## 1 INTRODUCTION

Document parsing aims to convert scanned documents into structured, machine-readable formats and represents one of the core tasks in document intelligence (Hwang et al., 2021; Wang et al., 2024b; Wei et al., 2024; Xia et al., 2024; Zhang et al., 2024). Unlike traditional OCR that focuses solely on text recognition, document parsing requires comprehensive recovery of hierarchical document structures, including the dependency relationships among elements such as paragraphs, headers, tables, and formulas—a capability that is crucial for downstream applications including legal contract analysis, scientific literature mining, and financial report processing. Traditional approaches typically rely on multi-stage pipelines that decompose the task into supervised sub-tasks—such as layout detection, OCR, table recognition, and formula recognition—followed by heuristic post-processing to reconstruct document structure (Blecher et al., 2024; Wei et al., 2025; Liu et al., 2024b; Wei et al., 2024; Bai et al., 2024; Chen et al., 2024). However, such pipeline-based methods are prone to error propagation and exhibit limited adaptability when confronted with diverse layout variations.

Recent approaches primarily reformulate document parsing as end-to-end perception tasks using vision-language models (VLMs) trained through supervised fine-tuning (SFT). However, this paradigm faces fundamental limitations. Although SFT provides token-level supervision, it often overfits to surface patterns rather than learning generalizable structural representations. This limitation is further compounded by the scarcity of large-scale, high-quality training data for document parsing, which hinders models from acquiring layout-aware knowledge. To overcome these limitations, reinforcement learning (RL) presents a promising alternative, having demonstrated strong generalization capabilities in vision and multimodal tasks, where outcome-based rewards help models learn transferable representations (Huang et al., 2025; Liu et al., 2025a; Wang et al., 2025c). This raises a fundamental question: can reinforcement learning drive models toward generalizable layout parsing rules? Unfortunately, current RL approaches remain limited by coarse-grained, binary outcome rewards that fail to provide the fine-grained, layout-aware supervision necessary for modeling

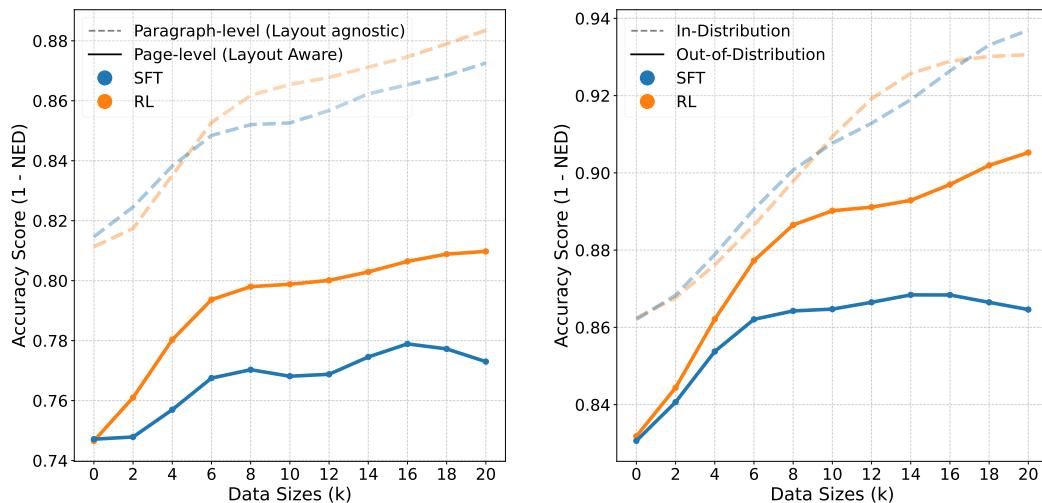


Figure 1: Comparison of document parsing performance on OmniDocBench under different training strategies as training data size increases. Left: Evaluation with two complementary metrics: (1) Paragraph-level accuracy (edit distance evaluation on element contents only), which assesses element-wise consistency within individual element contents, independent of inter-element reading order; and (2) Page-level accuracy (edit distance evaluation on element contents and reading order), which measures global document reconstruction quality by aligning predicted outputs (e.g., texts, tables, and formulas) with ground-truth sequences. Right: In-Distribution and Out-of-Distribution task performance measured by accuracy score (1 – NED). See detailed descriptions of the task in Section 4.3.

complex document layouts (Guo et al., 2025; Shao et al., 2024). Therefore, how to effectively apply RL to document parsing remains an underexplored yet critical challenge.

To address the challenge of effectively applying RL algorithms to document parsing tasks and the lack of large-scale, high-quality data in industry to support this training process, we propose LayoutRL, the first end-to-end reinforcement learning framework for layout-aware document parsing, and construct Infinity-Doc-400K, a large-scale dataset. Specifically, our approach does not rely on explicit reasoning processes, but rather treats the entire document parsing result as the final answer and guides model learning through carefully designed reward mechanisms. We introduce verifiable rewards (Shen et al., 2025; Liu et al., 2024a; Shao et al., 2024; Yaowei Zheng, 2025), which consist of Edit Distance Reward (Levenshtein et al., 1966) and Layout Parsing Reward signals, enforcing fine-grained alignment between predictions and ground-truth layouts and providing document-level supervision beyond token-level signals to encourage the model to learn transferable structural representations. Additionally, we construct a 400K-document corpus providing large-scale, high-quality supervision. It combines (1) high-fidelity synthetic scanned document parsing data—generated via HTML templates and browser rendering—and (2) expert-filtered real-world samples, pseudo-labeled through a cross-model agreement pipeline to capture genuine layout diversity. Built on this dataset, we train Infinity-Parser, an end-to-end VLM-based parser that directly outputs structured document representations.

To demonstrate the effectiveness of our approach, we conduct extensive experiments across multiple document parsing benchmarks. Results in Figure 1 show that while SFT achieves strong paragraph-level accuracy, its page-level performance struggles to improve as data size increases, reflecting its tendency to memorize surface patterns rather than model the hierarchical dependencies between document structures and elements. In contrast, our layout-aware reinforcement learning method significantly outperforms SFT in page-level accuracy, while maintaining competitive paragraph-level performance. These findings demonstrate that verifiable, layout-aware rewards enable models to move beyond simple token imitation, achieving consistent improvements in both local detail fidelity and global structural understanding, thereby establishing a more robust paradigm for document parsing. Moreover, our method also exhibits strong generalization to unseen task types in downstream applications. Finally, our method, Infinity-Parser, achieves new state-of-the-art results on four diverse benchmarks—OmniDocBench, olmOCR, PubTabNet, and FinTabNet—highlighting its effectiveness and strong cross-domain generalization.

108 We make the following contributions:  
 109

110 • We propose LayoutRL, a new reinforcement learning framework for end-to-end scanned docu-  
 111 ment parsing, which explicitly trains models to be layout-aware by optimizing verifiable, multi-  
 112 aspect rewards. Our multi-aspect reward design combines normalized edit distance, paragraph  
 113 count accuracy, and reading order preservation, improving structural robustness.  
 114 • We introduce Infinity-Doc-400K, a large-scale dataset of 400,482 scanned documents that com-  
 115 bines high-quality synthetic data with diverse real-world samples. The dataset features rich layout  
 116 variations and comprehensive structural annotations, enabling robust training.  
 117 • We train a VLM based model, Infinity-Parser, which sets new state-of-the-art performance across  
 118 English and Chinese benchmarks for OCR, table and formula extraction, and reading-order de-  
 119 tection—demonstrating substantial gains in both structural fidelity and semantic accuracy over  
 120 specialist pipelines and general-purpose vision-language models.  
 121

122 **2 RELATED WORK**  
 123

124 **2.1 REINFORCEMENT LEARNING FOR LANGUAGE MODELS**  
 125

126 Recent advancements in Large Language Models (LLMs) such as OpenAI’s GPT series (OpenAI,  
 127 2024)), DeepSeek-R1 (Guo et al., 2025), and Gemini (Team et al., 2023) have highlighted the sig-  
 128 nificant potential of Reinforcement Learning (RL) in enhancing their reasoning capabilities. This  
 129 RL paradigm has been successfully extended to other domains demanding sophisticated reasoning,  
 130 including code generation (Li et al., 2022; Zeng et al., 2025), autonomous tool utilization (Schick  
 131 et al., 2023; Wang et al., 2025a), and information retrieval (Nakano et al., 2021). Similarly, RL  
 132 has demonstrated its efficacy in the domain of Visual Language Models (VLMs), including pre-  
 133 cise object counting (Peng et al., 2025), nuanced visual perception (Liu et al., 2025b), and complex  
 134 multimodal reasoning (e.g., VL-Rethinker (Wang et al., 2025b), Pixel Reasoner (Su et al., 2025),  
 135 Vision-R1 (Huang et al., 2025)). These pioneering works have predominantly relied on binary out-  
 136 come rewards to guide RL training. Complementary to these efforts, our work demonstrates the  
 137 effectiveness of incorporating layout-aware and layout-based rewards for document parsing, offer-  
 138 ing a more granular and contextually relevant feedback mechanism.  
 139

140 **2.2 VLM-BASED DOCUMENT PARSING**  
 141

142 Recent advancements in document understanding and optical character recognition (OCR) have  
 143 highlighted their importance as critical benchmarks for evaluating the perceptual capabilities of  
 144 vision-language models (VLMs). By incorporating large-scale OCR corpora during pretraining,  
 145 models such as GPT-4o (Achiam et al., 2023) and Qwen2-VL (Bai et al., 2024) have achieved  
 146 competitive performance on document content extraction tasks. Building upon these foundations,  
 147 the emergence of VLMs has further accelerated the progress of end-to-end document parsing, giving  
 148 rise to a range of models such as Donut (Blecher et al., 2024), Nougat (Blecher et al., 2023), Kosmos-  
 149 2.5 (Lv et al., 2024), Vary (Wei et al., 2025), mPLUG-DocOwl (Hu et al., 2024b), Fox (Liu et al.,  
 150 2024b), and GOT (Wei et al., 2024). These models have continued to improve their understanding  
 151 of visual layouts and textual content by leveraging advancements in visual encoders (Dosovitskiy  
 152 et al., 2020), language decoders (Bai et al., 2024), and data construction pipelines. Despite the  
 153 success of these VLM-based approaches in enabling end-to-end document parsing, they still face  
 154 generalization challenges on downstream layout parsing tasks (Wang et al., 2024b). To address this  
 155 issue, we propose leveraging reinforcement learning to provide a more effective training paradigm  
 156 that better aligns with the demands of document parsing.  
 157

158 **3 METHODOLOGY**  
 159

160 In this section, we first introduce Infinity-Doc-400K, our large-scale multimodal dataset for end-to-  
 161 end scanned document parsing. We then describe our rule-based multi-aspect reward framework,  
 162 which integrates edit distance, paragraph count, and order criteria under a unified reinforcement  
 163 learning objective optimized via Group Relative Policy Optimization (GRPO).  
 164

Benchmark	Document Domain	Annotation Type					End-to-End Task				Exactly Match
		BBox	Text	Table	Formula	Attributes	OCR	TR	MFR	ROD	
<b>End-to-end Eval Benchmarks</b>											
Fox (Liu et al., 2024b)	2	✓	✓				✓				
Nougat (Blecher et al., 2024)	1		✓	✓	✓		✓	✓	✓		
GOT OCR 2.0 (Wei et al., 2024)	2		✓	✓	✓	✓	✓	✓	✓		
OmniDocBench (Ouyang et al., 2024)	9	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<b>End-to-end Train Dataset</b>											
DocStruct4M (Hu et al., 2024a)	-		✓				✓				
olmOCR-mix (Poznanski et al., 2025)	-		✓	✓	✓	✓	✓	✓	✓	✓	
<b>Infinity-Doc-400K</b>	7	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: A comparison between Infinity-Doc-400K and existing datasets. *BBox*: Bounding boxes. *Text*: Text in Unicode. *Table*: Table in LaTeX/HTML/Markdown. *Formula*: Formula in LaTeX. *Attributes*: Page- and BBox-Level Attributes. *OCR*: Optical Character Recognition; *TR*: Table Recognition; *MFR*: Math Formula Recognition; *ROD*: Reading Order Detection. *Multi-Type Doc*: Whether the dataset includes documents from multiple domains or categories.

### 3.1 INFINITY-DOC-400K AND GENERATION PIPELINES

We introduce Infinity-Doc-400K, a large-scale, multimodal dataset of 400,066 richly annotated documents for end-to-end scanned document parsing. Unlike prior benchmarks that target isolated subtasks (e.g., layout detection, OCR, or table recognition), Infinity-Doc-400K provides holistic supervision by pairing rendered scanned document pages with their ground-truth Markdown representations. This design enables training and evaluating models that directly translate visual inputs to layout outputs without relying on brittle, multi-stage pipelines. As shown in Table 1, compared to existing works, Infinity-Doc-400K not only significantly enhances task diversity but also substantially improves overall data quality through our proposed synthetic generation mechanism. More details on the data distribution and quality control are provided in the Data Details section of the Appendix.

To construct Infinity-Doc-400K, we design a dual-pipeline framework that integrates both synthetic and real-world document generation, as illustrated in Figure 2. This design addresses a critical limitation of traditional data construction pipelines, which often rely on weak supervision and pseudo-labeling from a single model applied to crawled, scanned documents. These pipelines frequently suffer from noisy, misaligned, or incomplete annotations, especially in complex layouts or multilingual content, thus hindering model performance and generalization. To overcome these issues, our dual-pipeline framework is motivated by the need to balance annotation quality and structural diversity. The synthetic branch provides highly accurate, clean, and precisely aligned annotations at scale, while the real-world branch introduces naturally occurring layout variability and semantic richness, which are essential for building models that generalize robustly in practical applications.

**Real-World Data** We develop a real-world data construction pipeline to capture the structural complexity and natural layout variability of documents across practical domains. We collect diverse scanned documents from sources such as financial reports, medical records, academic papers, books, magazines, and web pages, covering both dense and sparse content layouts. To generate structural annotations, we adopt a multi-expert strategy, where specialized models handle different structural elements, such as layout blocks, texts, formulas, and tables. For example, overall layouts are analyzed by a visual layout model (Huang et al., 2022), formula regions are processed by a dedicated formula recognition model (Wang et al., 2024a), and tables are parsed by a transformer-based table extractor (Blecher et al., 2024). A cross-validation mechanism is then applied to filter out inconsistencies by comparing the outputs of expert models and VLMs. Only regions with consistent predictions across models are retained as high-confidence pseudo-ground-truth annotations. This layout-aware filtering results in a rich and reliable dataset that reflects the complexity of real-world documents and supports robust document parsing model training.

**Synthetic Data** We design a synthetic data construction pipeline. We collect text and images from sources such as Wikipedia, web crawlers, and online corpora, and use Jinja (Nipkow, 2003) templates to inject sampled content into predefined single-, double-, or triple-column HTML layouts. These pages are rendered into scanned documents using a browser engine, followed by automated filtering to remove low-quality or overlapping images. Ground-truth annotations are extracted by

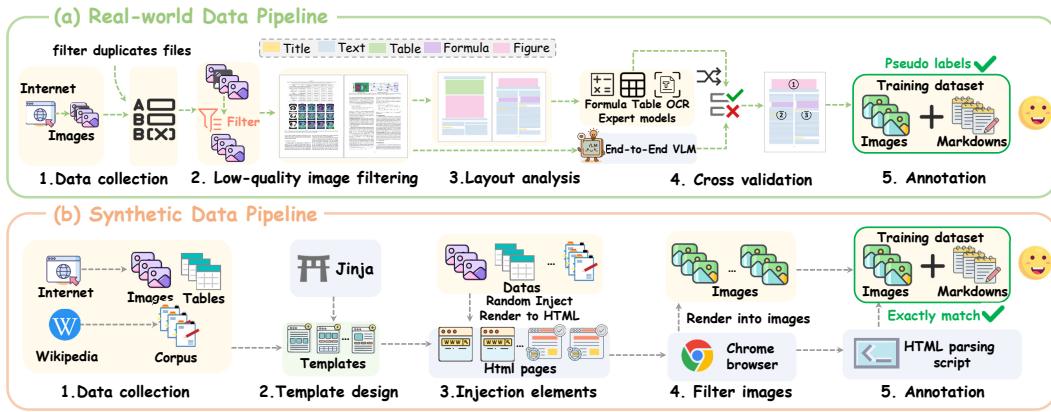


Figure 2: Data construction pipelines for document parsing. (a) Real-world pipelines enhance quality by combining multiple expert models and layout analysis, yielding better-aligned supervision through intersection and reading order reasoning. (b) Synthetic pipeline leverages structured HTML templates and browser rendering to generate clean, exactly-aligned scanned document parsing data, ensuring high-quality supervision for end-to-end parsing.

parsing the original HTML to produce aligned Markdown representations. This synthetic approach not only significantly reduces construction costs and ensures annotation accuracy and structural diversity, but more importantly, it addresses the longstanding issue of imprecise or inconsistent supervision commonly found in pseudo-labeled datasets, providing high-quality and well-aligned supervision for training end-to-end models.

### 3.2 RL WITH LAYOUT-AWARE REWARDS

As illustrated in Figure 3, we employ an RL framework to directly optimize scanned document parsers, aiming to enhance both structural fidelity and semantic accuracy. Specifically, we utilize GRPO (Shao et al., 2024), which enables learning from rule-based reward signals without relying on absolute values. GRPO operates by generating a set of candidate Markdown outputs for each document and evaluating them using a *multi-aspect reward*, denoted as  $R_{\text{Multi-Aspect}}$ , which integrates multiple rule-based criteria into a unified supervisory signal. These raw rewards are then converted into relative advantage scores by comparing each candidate against others within the same group. This relative evaluation promotes training stability and encourages the selection of higher-quality outputs, eliminating the need for a learned value function or critic. Notably, our reinforcement learning approach avoids any explicit thinking or intermediate reasoning process; instead, all outputs are treated as final answers, with the model receiving verifiable rewards based on these outputs.

The multi-aspect reward  $R_{\text{Multi-Aspect}}$  consists of three complementary components, each capturing a different aspect of parsing quality:

**Edit Distance Reward ( $R_{\text{dist}}$ )** We define the edit distance reward based on the normalized Levenshtein distance  $D(y, \hat{y})$  between the predicted output  $\hat{y}$  and the reference output  $y$ :

$$R_{\text{dist}} = 1 - \frac{D(y, \hat{y})}{\max(N, M)} \quad (1)$$

where  $N = |y|$  and  $M = |\hat{y}|$  are the lengths of the reference and predicted sequences, respectively. The distance  $D(y, \hat{y})$  measures the minimum number of single-character insertions, deletions, or substitutions required to convert  $\hat{y}$  into  $y$ , thereby capturing both semantic and formatting discrepancies. This reward is bounded within  $[0, 1]$ , with higher values indicating better alignment between prediction and reference. Meanwhile, the reference output  $y$  is synthesized through two data generation pipelines proposed in this work. It is produced with rigorous rule-based filtering and consistency validation, and serves as a high-quality surrogate for ground-truth annotations in evaluating the quality of the model output.

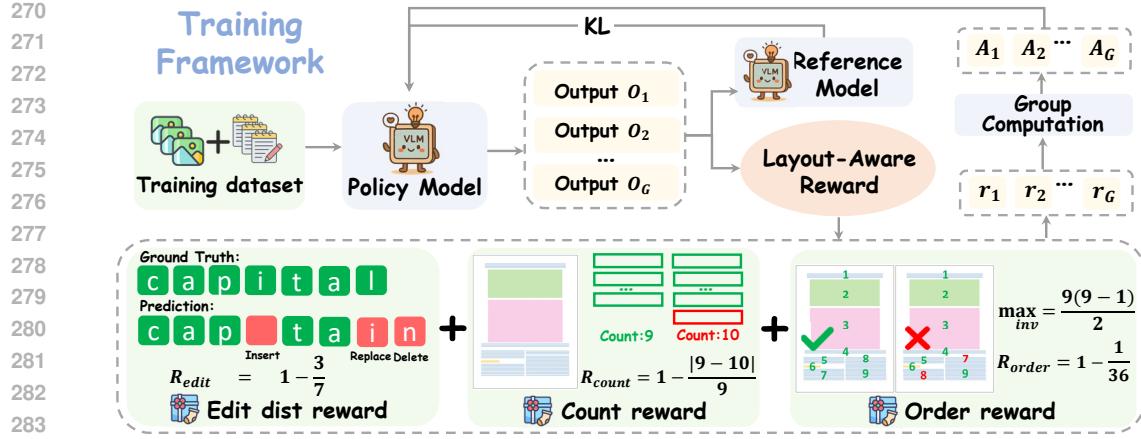


Figure 3: Overview of Infinity-Parser training framework. Our model is optimized via reinforcement finetuning with edit distance, layout, and order-based rewards.

**Count Reward ( $R_{\text{count}}$ )** To encourage accurate paragraph segmentation, let  $N_Y$  and  $N_{\hat{Y}}$  be the numbers of reference and predicted paragraphs. We define:

$$R_{\text{count}} = 1 - \frac{|N_Y - N_{\hat{Y}}|}{N_Y} \quad (2)$$

which penalizes missing or spurious paragraphs.

**Order Reward ( $R_{\text{order}}$ )** We measure sequence-level fidelity by counting pairwise inversions  $D_{\text{order}}$  between reference and predicted paragraphs. with  $\text{max}_{\text{inv}} = N_Y(N_Y - 1)/2$ , we set:

$$R_{\text{order}} = 1 - \frac{D_{\text{order}}}{\text{max}_{\text{inv}}} \quad (3)$$

rewarding preservation of the original reading order.

The final multi-aspect reward is a weighted combination of these three components. Specifically, we begin by applying the Hungarian algorithm (Kuhn, 1955) to establish the optimal one-to-one matching between predicted and ground-truth segments, identifying both pairings and their relative order. Based on the matched segment count, we compute the count reward to reflect alignment in the number of segments. Using the relative sequence of matched pairs, we calculate the order reward to measure structural consistency. On top of these matchings, we compute the edit reward by averaging the edit similarities of each matched segment pair. Combining these terms yields the final reward:

$$R_{\text{Multi-Aspect}} = R_{\text{dist}} + R_{\text{count}} + R_{\text{order}} \quad (4)$$

This multi-aspect design balances content fidelity with structural correctness and order preservation, providing rich supervision for end-to-end document parsing.

## 4 EXPERIMENTS

We adopt Qwen2.5-VL-7B (Bai et al., 2025a) as the base model and apply the VeRL (Sheng et al., 2024) framework for reinforcement learning. Detailed implementation can be found in the Appendix.

### 4.1 MAIN RESULTS

We evaluate our method on several widely-used benchmarks for document understanding and OCR tasks. OmniDocBench (Ouyang et al., 2024) provides comprehensive evaluation across diverse document types using NED and TEDS metrics. For table recognition, we use PubTabNet with scientific tables and FinTabNet (Zheng et al., 2021) with financial documents. Additionally, we

Methods	Overall <sup>Edit↓</sup>		Text <sup>Edit↓</sup>		Form. <sup>Edit↓</sup>		Table <sup>TEDS↑</sup>		Table <sup>Edit↓</sup>		Read Order <sup>Edit↓</sup>	
	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH
<b>Based on Pipeline Tools</b>												
MinerU (Wang et al., 2024b)	0.15	0.357	<b>0.061</b>	0.215	<b>0.278</b>	0.577	78.6	62.1	0.18	0.344	0.079	0.292
Marker (Paruchuri, 2024)	0.336	0.556	0.080	0.315	0.530	0.883	67.6	49.2	0.619	0.685	0.114	0.340
Mathpix	0.191	0.365	0.105	0.384	0.306	0.454	77.0	67.1	0.243	0.320	0.108	0.304
Docling (Livathinos et al., 2025)	0.589	0.909	0.416	0.987	0.999	1.000	61.3	25.0	0.627	0.810	0.313	0.837
Pix2Text (Gurgurov & Morshnev, 2024)	0.320	0.528	0.138	0.356	0.276	0.611	73.6	66.2	0.584	0.645	0.281	0.499
Unstructured-0.17.2	0.586	0.716	0.198	0.481	0.999	1.000	-	-	1.000	0.998	0.145	0.387
OpenParse-0.7.0	0.646	0.814	0.681	0.974	0.996	1.000	64.8	27.5	0.284	0.639	0.595	0.641
<b>Based on Expert VLMs</b>												
GOT-OCR (Wei et al., 2024)	0.287	0.411	0.189	0.315	0.360	0.528	53.2	47.2	0.459	0.520	0.141	0.280
Nougat (Blecher et al., 2024)	0.452	0.973	0.365	0.998	0.488	0.941	39.9	0.0	0.572	1.000	0.382	0.954
Mistral OCR	0.268	0.439	0.072	0.325	0.318	0.495	75.8	63.6	0.600	0.650	0.083	0.284
OLMOCR-slang	0.326	0.469	0.097	0.293	0.455	0.655	68.1	61.3	0.608	0.652	0.145	0.277
SmolDocling-256M	0.493	0.816	0.262	0.838	0.753	0.997	44.9	16.5	0.729	0.907	0.227	0.522
<b>Based on General VLMS</b>												
GPT-4o (Achiam et al., 2023)	0.233	0.399	0.144	0.409	0.425	0.606	72.0	62.9	0.234	0.329	0.128	0.251
Qwen2-VL-72B (Wang et al., 2024c)	0.252	0.327	0.096	0.218	0.404	0.487	76.8	76.4	0.387	0.408	0.119	0.193
InternVL2-76B (Chen et al., 2024)	0.440	0.443	0.353	0.290	0.543	0.701	63.0	60.2	0.547	0.555	0.317	0.228
Qwen2.5-VL-7B (Bai et al., 2025b)	0.220	0.265	0.142	0.205	0.393	0.530	78.7	78.3	0.155	0.162	0.191	0.169
InternVL3-8B (Zhu et al., 2025)	0.426	0.385	0.315	0.345	0.714	0.729	59.0	71.5	0.352	0.211	0.324	0.257
<b>Based on Reinforcement Learning</b>												
Infinity-Parser-7B	<b>0.141</b>	<b>0.197</b>	0.076	<b>0.117</b>	0.314	<b>0.434</b>	<b>85.3</b>	<b>81.4</b>	<b>0.098</b>	<b>0.142</b>	<b>0.076</b>	<b>0.095</b>

Table 2: Comprehensive evaluation of document parsing algorithms on OmniDocBench: performance metrics for text, formula, table, and reading order extraction, with overall scores derived from ground truth comparisons.

employ olmOCR-Bench (Poznanski et al., 2025) for fact-based OCR evaluation. We ensure that the test data for each benchmark undergoes rigorous text similarity filtering to prevent any overlap with the training data. Detailed descriptions of these benchmarks are provided in Appendix.

**Overall Evaluation on OmniDocBench** As shown in Table 2, pipeline-based methods such as MinerU (Wang et al., 2024b) and Mathpix achieve superior performance across individual sub-tasks including text recognition and formula recognition. Meanwhile, general-purpose vision-language models like Qwen2.5-VL-7B and GPT-4o also demonstrate competitive results. Notably, most methods perform better on English pages compared to Chinese pages, reflecting language-dependent challenges. In contrast, our proposed Infinity-Parser-7B achieves a more balanced performance across all sub-tasks and languages, setting new SOTA results with overall edit distances of 0.141 and 0.197. This highlights the advantage of reinforcement learning with multi-aspect rewards in enabling robust, end-to-end document parsing.

Model	Overall	ArXiv	Old Scans Math	Tables	Old Scans	Headers&Footers	Multi Col.	Long-Tiny Text	Base
GOT OCR	48.3	52.7	52.0	0.2	22.1	93.6	42.0	29.9	94.0
Marker v1.6.2	59.4	24.3	22.1	69.8	24.3	87.1	71.0	76.9	99.5
MinerU v1.3.10	61.5	75.4	47.4	60.9	17.3	<b>96.6</b>	59.0	39.1	96.6
Mistral OCR API	72.0	77.2	67.5	60.6	29.3	93.6	71.3	77.1	99.4
GPT-4o (No Anchor)	68.9	51.5	75.5	69.1	40.9	94.2	68.9	54.1	96.7
GPT-4o (Anchored)	69.9	53.5	74.5	70.0	40.7	93.8	69.3	60.6	96.8
Gemini Flash 2 (No Anchor)	57.8	32.1	56.3	61.4	27.8	48.0	58.7	84.4	94.0
Gemini Flash 2 (Anchored)	63.8	54.5	56.1	72.1	34.2	64.7	61.5	71.5	95.6
Qwen 2 VL (No Anchor)	31.5	19.7	31.7	24.2	17.1	88.9	8.3	6.8	55.5
Qwen 2.5 VL (No Anchor)	65.5	63.1	65.7	67.3	38.6	73.6	68.3	49.1	98.3
olmOCR v0.1.68 (No Anchor)	76.3	72.1	74.7	71.5	43.7	91.6	78.5	80.5	98.1
olmOCR v0.1.68 (Anchored)	77.4	75.6	75.1	70.2	44.5	93.4	79.4	81.7	99.0
Infinity-Parser-7B	<b>82.5</b>	<b>84.4</b>	<b>83.8</b>	<b>85.0</b>	<b>47.9</b>	88.7	<b>84.2</b>	<b>86.4</b>	<b>99.8</b>

Table 3: Performance comparison on the olmOCR (Poznanski et al., 2025) benchmark across multiple document domains and structural challenges. Higher is better.

**Document-level OCR Evaluation** Table 3 reports performance on the olmOCR-Bench benchmark, which evaluates document-level OCR across diverse layouts and domains. Infinity-Parser-7B achieves the highest overall score (82.5), followed closely by olmOCR v0.1.68 (Anchored) (77.4), both demonstrating strong performance in complex categories like multi-column layouts and scanned math content. The results highlight the effectiveness of anchored prompting, with anchored versions of models (e.g., GPT-4o, olmOCR) significantly outperforming their non-anchored counterparts—especially on tables and old scans. This underscores the importance of layout-aware extraction techniques. In contrast, traditional pipelines like Marker and GOT OCR lag behind in

378 structural accuracy, reinforcing the value of modern VLM-based approaches in high-fidelity PDF  
 379 understanding.  
 380

381 **Table Recognition Evaluation** To evaluate  
 382 the model’s generalization ability, we intro-  
 383 duce task-specific test cases. In Table 4,  
 384 we compare Infinity-Parser-7B with end-to-  
 385 end table recognition models on PubTab-  
 386 Net and FinTabNet using the TEDS  
 387 metric, which evaluates both structure and  
 388 content. We also report TEDS-S for structure-  
 389 only assessment. The evaluation results for In-  
 390 nternVL3, Qwen2.5-VL, and GPT-4o were  
 391 generated through our standardized benchmarking  
 392 pipeline. Infinity-Parser-7B achieves the high-  
 393 est TEDS-S and TEDS scores on both datasets.  
 394

## 395 4.2 ABLATION STUDY

396 We perform ablation experiments to evaluate the individual contributions of our three core design  
 397 choices: (1) data quality verification and (2) multi-aspect rewards. We report all ablation results  
 398 using two primary evaluation metrics:  $\text{Overall}^{\text{Edit}}$  and  $\text{Overall}^{\text{Cat.}}$ .  $\text{Overall}^{\text{Edit}}$  represents the aver-  
 399 age edit-based overall score across English and Chinese pages, as shown in Table 2. In contrast,  
 400  $\text{Overall}^{\text{Cat.}}$  reflects the mean category-level performance across nine types of scanned document  
 401 pages, following the same evaluation setting as Table 9 in the Appendix.  
 402

Method	Edit Dist.	Count.	Order.	SFT	RL	$\text{Overall (EN)}^{\text{Edit} \downarrow}$	$\text{Overall (ZH)}^{\text{Edit} \downarrow}$	$\text{Overall}^{\text{Cat.} \downarrow}$
Zero Shot	-	-	-	-	-	0.220	0.265	0.183
SFT	-	-	-	43K	-	0.198	0.261	0.159
Zero + RL	✓	-	-	-	43K	0.169	0.224	0.156
Zero + RL	✓	✓	-	-	43K	0.159	0.200	0.112
Zero + RL	✓	✓	✓	-	43K	0.141	0.197	0.104
SFT + RL	✓	✓	✓	43K	43K	0.163	0.195	0.092

411 Table 5: Results under different reward designs.

412 **Effect of Multi-Aspect Rewards.** Table 5 demonstrates that reinforcement learning can outperform  
 413 supervised fine-tuning when appropriate reward designs are applied. Compared to the SFT baseline  
 414 (0.198 / 0.261 / 0.159), the RL method with distance-based reward ( $\text{Zero} + R_{\text{dist}}$ ) achieves better  
 415  $\text{Overall}^{\text{Edit}}$  (EN: 0.169 vs. 0.198, ZH: 0.224 vs. 0.261) while maintaining a comparable  $\text{Overall}^{\text{Cat.}}$   
 416 (0.156 vs. 0.159). Incorporating additional count and order rewards further improves structural  
 417 consistency:  $\text{Zero} + R_{\text{dist}} + R_{\text{count}}$  achieves 0.159 / 0.200 / 0.112, and  $\text{Zero} + R_{\text{dist}} + R_{\text{count}} +$   
 418  $R_{\text{order}}$  achieves 0.141 / 0.197 / 0.104. Meanwhile, when reinforcement learning is combined with  
 419 supervised fine-tuning ( $\text{SFT} + \text{RL}$ ), the model does not exhibit further significant improvements.  
 420 This observation is consistent with existing studies (Guo et al., 2025), which suggest that given  
 421 the backbone model’s inherent capabilities, SFT may be unnecessary for downstream RL training  
 422 in certain scenarios. These results further demonstrate that reinforcement learning, when equipped  
 423 with structural supervisory signals, enables the model to better align with task-specific objectives.  
 424

## 425 4.3 FURTHER ANALYSIS OF LAYOUTRL

426 **Training Stability Across Task Types.** As shown in Figure 4, we compare SFT and Layout-  
 427 Aware RL on four OmniDocBench sub-tasks: table recognition, text recognition, formula parsing,  
 428 and reading order prediction. Across all tasks, RL consistently achieves better performance, yielding  
 429 higher scores on TEDS and lower errors on NED. More importantly, the RL curves exhibit smoother  
 430 trajectories with stable improvements over training, while SFT shows large fluctuations and even  
 431 performance regressions at certain stages. These results indicate that layout-aware rewards not only  
 432 improve final accuracy but also enhance training stability throughout optimization.

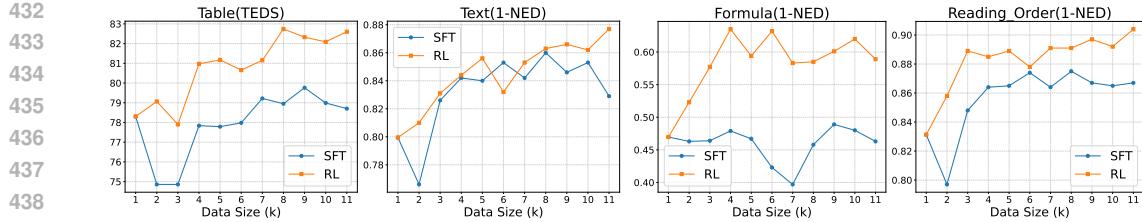


Figure 4: Performance comparison of SFT and Layout-Aware RL on OmniDocBench sub-tasks.

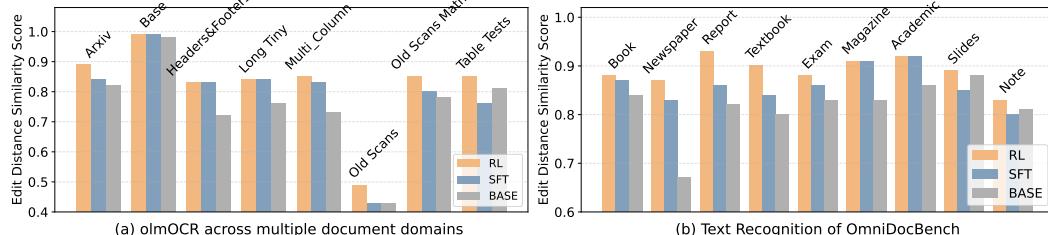


Figure 5: Comparison of model performance on different document parsing tasks.

**Robustness Across Diverse Document Tasks.** Figure 5 compares RL, SFT, and Zero-Shot (Base) across diverse document parsing tasks. On olmOCR (left), RL consistently achieves higher Levenshtein distance similarity score, especially on challenging cases like old scans and table tasks. At the same time, SFT offers moderate gains over Zero-Shot but remains behind RL. On OmniDocBench (right), RL also outperforms the other methods across most document types, showing notable improvements on books, reports, and academic texts. Overall, RL demonstrates greater robustness and better generalization in both structural parsing on olmOCR and text recognition on omnidocbench.

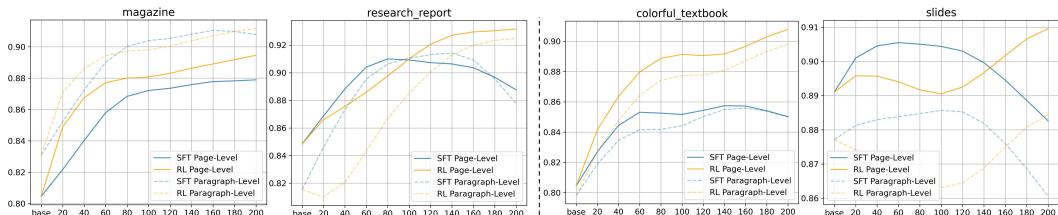


Figure 6: Generalization comparison of SFT and Layout-Aware RL. X-axis represents training steps. Y-axis represents 1-NED scores.

**Analysis of Generalization** The Figure 6 compares document parsing performance across different training steps. The left two plots (magazine, research report) correspond to In-Distribution settings, where training and evaluation domains are aligned, while the right two plots (colorful textbook, slides) correspond to OOD settings, where evaluation involves unseen document types. In the in-distribution case, SFT achieves stable paragraph-level accuracy but its page-level performance tends to plateau, reflecting reliance on surface patterns. By contrast, RL continues to improve with data scale, achieving notable gains in page-level accuracy. In the OOD case, SFT performance degrades more severely, while RL maintains robustness and shows stronger improvements, highlighting its ability to capture global structural dependencies and generalize across distributions.

## 5 CONCLUSION

We introduced LayoutRL, an end-to-end reinforcement learning framework that explicitly incorporates layout awareness into document parsing through verifiable, multi-aspect rewards. To support this training, we built Infinity-Doc-400K, a large-scale dataset combining synthetic and real-world documents with diverse layouts, and trained Infinity-Parser, a VLM-based parser. Experiments on OmniDocBench, olmOCR-Bench, PubTabNet, and FinTabNet show that our approach achieves state-of-the-art performance across languages and document types, outperforming both specialized pipelines and general-purpose VLMs. Beyond accuracy, LayoutRL improves training stability and demonstrates robustness across diverse document tasks, highlighting reinforcement learning as a promising direction for robust and transferable document intelligence.

486 REFERENCES  
487

488 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Ale-  
489 man, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical  
490 report. *arXiv:2303.08774*, 2023.

491 Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang  
492 Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, local-  
493 ization, text reading, and beyond. *arXiv:2308.12966*, 2024.

494

495 Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang,  
496 Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*,  
497 2025a.

498 Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang,  
499 Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*,  
500 2025b.

501 Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical  
502 understanding for academic documents. *arXiv preprint arXiv:2308.13418*, 2023.

503 Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical  
504 understanding for academic documents. *arXiv:2308.13418*, 2024.

505

506 Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong  
507 Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl:  
508 Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Pro-  
509 ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–  
510 24198, June 2024.

511

512 Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas  
513 Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An  
514 image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint  
515 arXiv:2010.11929*, 2020.

516

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

520 Daniil Gurgurov and Aleksey Morshnev. Image-to-latex converter for mathematical formulas and  
521 text. *arXiv preprint arXiv:2408.04015*, 2024.

522

523 Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin,  
524 Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document under-  
525 standing. *arXiv preprint arXiv:2403.12895*, 2024a.

526

527 Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and  
528 Jingren Zhou. mplug-docowl2: High-resolution compressing for ocr-free multi-page document  
529 understanding. *arXiv preprint arXiv:2409.03420*, 2024b.

530

531 Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Yao Hu, and Shaohui  
532 Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv  
533 preprint arXiv:2503.06749*, 2025.

534

535 Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for  
536 document ai with unified text and image masking, 2022. URL <https://arxiv.org/abs/2204.08387>.

537

538 Wonseok Hwang, Jinyeong Yim, Seunghyun Park, Sohee Yang, and Minjoon Seo. Spatial depen-  
539 dency parsing for semi-structured document information extraction. In *Findings of the Associa-  
540 tion for Computational Linguistics: ACL-IJCNLP*, pp. 330–343. Association for Computational  
541 Linguistics (ACL), 2021.

540 Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics*  
 541 *quarterly*, 2(1-2):83–97, 1955.

542

543 Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals.  
 544 In *Doklady Physics*, volume 10, pp. 707–710. Soviet Union, 1966.

545

546 Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittweis, Rémi Leblond, Tom  
 547 Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. Competition-level code generation  
 548 with alphacode. *Science*, 378(6624):1092–1097, 2022.

549

550 Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao,  
 551 Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint*  
 551 *arXiv:2412.19437*, 2024a.

552

553 Chenglong Liu, Haoran Wei, Jinyue Chen, Lingyu Kong, Zheng Ge, Zining Zhu, Liang Zhao, Jian-  
 554 jian Sun, Chunrui Han, and Xiangyu Zhang. Focus anywhere for fine-grained multi-page docu-  
 555 ment understanding. *arXiv:2405.14295*, 2024b.

556

557 Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Vision-  
 558 reasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv preprint*  
 558 *arXiv:2505.12081*, 2025a.

559

560 Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi  
 561 Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025b.

562

563 Nikolaos Livathinos, Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Panos Vage-  
 564 nas, Cesar Berrospi Ramis, Matteo Omenetti, Kasper Dinkla, Yusik Kim, et al. Docling: An ef-  
 565 ficient open-source toolkit for ai-driven document conversion. *arXiv preprint arXiv:2501.17887*,  
 566 2025.

567

568 Tengchao Lv, Yupan Huang, Jingye Chen, Yuzhong Zhao, Yilin Jia, Lei Cui, Shuming Ma, Yaoyao  
 569 Chang, Shaohan Huang, Wenhui Wang, Li Dong, Weiyao Luo, Shaoxiang Wu, Guoxin Wang,  
 570 Cha Zhang, and Furu Wei. Kosmos-2.5: A multimodal literate model, 2024. URL <https://arxiv.org/abs/2309.11419>.

571

572 Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christo-  
 573 pher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted  
 574 question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

575

576 Tobias Nipkow. Jinja: Towards a comprehensive formal semantics for a java-like language. In *Proc.  
 577 Marktoberdorf Summer School*. IOS Press Amsterdam, 2003.

578

579 OpenAI. Gpt-4o system card, 2024. URL <https://arxiv.org/abs/2410.21276>.

580

581 Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan  
 582 Zhao, Man Jiang, Xiaomeng Zhao, et al. Omnidocbench: Benchmarking diverse pdf document  
 583 parsing with comprehensive annotations. *arXiv preprint arXiv:2412.07626*, 2024.

584

585 Vik Paruchuri. Marker, 2024. URL <https://github.com/VikParuchuri/marker>.

586

587 Yi Peng, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun  
 588 Pan, Tianyidan Xie, Li Ge, et al. Skywork r1v: pioneering multimodal reasoning with chain-of-  
 589 thought. *arXiv preprint arXiv:2504.05599*, 2025.

590

591 Jake Poznanski, Jon Borchardt, Jason Dunkelberger, Regan Huff, Daniel Lin, Aman Rangapur,  
 592 Christopher Wilhelm, Kyle Lo, and Luca Soldaini. olmocr: Unlocking trillions of tokens in  
 593 pdfs with vision language models. *arXiv preprint arXiv:2502.18443*, 2025.

594

595 Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro,  
 596 Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can  
 597 teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–  
 598 68551, 2023.

594 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,  
 595 Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical  
 596 reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

597

598 Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun  
 599 Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large  
 600 vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

601 Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng,  
 602 Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint*  
 603 *arXiv: 2409.19256*, 2024.

604

605 Alex Su, Haozhe Wang, Weimin Ren, Fangzhen Lin, and Wenhui Chen. Pixel reasoner: In-  
 606 centivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint*  
 607 *arXiv:2505.15966*, 2025.

608 Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut,  
 609 Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly  
 610 capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

611

612 Bin Wang, Zhuangcheng Gu, Guang Liang, Chao Xu, Bo Zhang, Botian Shi, and Conghui He.  
 613 Unimernet: A universal network for real-world mathematical expression recognition, 2024a. URL  
 614 <https://arxiv.org/abs/2404.15254>.

615

616 Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu,  
 617 Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua  
 618 Lin, and Conghui He. Mineru: An open-source solution for precise document content extraction.  
*arXiv:2409.18839*, 2024b.

619

620 Haozhe Wang, Long Li, Chao Qu, Fengming Zhu, Weidi Xu, Wei Chu, and Fangzhen Lin. To code  
 621 or not to code? adaptive tool integration for math language models via expectation-maximization.  
*arXiv preprint arXiv:2502.00691*, 2025a.

622

623 Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. VI-  
 624 rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning.  
*arXiv preprint arXiv:2504.08837*, 2025b.

625

626 Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. VI-  
 627 rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning.  
*arXiv preprint arXiv:2504.08837*, 2025c.

628

629 Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu,  
 630 Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the  
 631 world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024c.

632

633 Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang  
 634 Zhao, Jianjian Sun, Yuang Peng, et al. General ocr theory: Towards ocr-2.0 via a unified end-to-  
 635 end model. *arXiv:2409.01704*, 2024.

636

637 Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chun-  
 638 rui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language  
 639 model. In *European Conference on Computer Vision*, pp. 408–424. Springer, 2025.

640

641 Renqiu Xia, Song Mao, Xiangchao Yan, Hongbin Zhou, Bo Zhang, Haoyang Peng, Jiahao Pi,  
 642 Daocheng Fu, Wenjie Wu, Hancheng Ye, et al. Docgenome: An open large-scale scientific doc-  
 643 ument benchmark for training and testing multi-modal large language models. *arXiv preprint*  
*arXiv:2406.11633*, 2024.

644

645 Junting Lu Yaowei Zheng. Easyr1: An efficient, scalable, multi-modality rl training framework.  
 646 <https://github.com/hiyouga/EasyR1>, 2025.

647

648 Huaye Zeng, Dongfu Jiang, Haozhe Wang, Ping Nie, Xiaotong Chen, and Wenhui Chen. Acecoder:  
 649 Acing coder rl via automated test-case synthesis. *arXiv preprint arXiv:2502.01718*, 2025.

648 Qintong Zhang, Victor Shea-Jay Huang, Bin Wang, Junyuan Zhang, Zhengren Wang, Hao Liang,  
649 Shawn Wang, Matthieu Lin, Wentao Zhang, and Conghui He. Document parsing unveiled:  
650 Techniques, challenges, and prospects for structured information extraction. *arXiv preprint*  
651 *arXiv:2410.21169*, 2024.

652 Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. Global table  
653 extractor (gte): A framework for joint table identification and cell structure recognition using  
654 visual context. In *Proceedings of the IEEE/CVF winter conference on applications of computer*  
655 *vision*, pp. 697–706, 2021.

656 Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao  
657 Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for  
658 open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701