Table-MCR2TR: Merged-Cell-Aware Table Recognition via Reinforced Multimodal Language Models

Bangbang Zhou[1⋆], Zhaoqing Zhu[2], Feiyu Gao[2], Hangdi Xing[2], Yadong Qu[1], Qi Zheng[2], Ming Yan[2], and Hongtao Xie[1†]

[1] University of Science and Technology of China
[2] Tongyi Lab, Alibaba Group
{bangzhou01,qqqyd}@mail.ustc.edu.cn, htxie@ustc.edu.cn
{zhuzhaoqing.zzq,hangdi.xhd,feiyu.gfy,ym119608}@alibaba-inc.com
{yongqi.zq}@taobao.com

Abstract. Table recognition (TR) aims to transcribe table images into semi-structured representations (*e.g.*, HTML or Markdown). However, current methods still struggle with complex structures, particularly involving merged cells, which are essential for accurate table parsing and downstream understanding tasks. While some works improve TR through large-scale training and global supervision, these local fine-grained yet crucial merged-cell attributes are often overlooked, becoming a key bottleneck for further progress. To tackle this, we introduce Table-MCR2TR, a reinforced multimodal large language model framework that leverages the enhanced merged-cell recognition (MCR) ability as contextual guidance to improve table recognition quality. To enable this capability, we first develop a high-fidelity data pipeline to construct a large-scale table dataset covering many complex merged cells, alleviating their scarcity in existing datasets. Additionally, the merged-cell-aware table recognition (MCATR) optimization strategy is proposed to improve table parsing via collaborative reinforcement learning. The core of MCATR lies in integrating three complementary objectives: TR and MCR respectively learn the global table structure and fine-grained merged-cell information; based on these learned signals, the MCR2TR alignment task transfers merged-cell knowledge to guide TR generation. Experimental results show that Table-MCR2TR achieves SOTA across multiple benchmarks, with average TR/MCR accuracy of 92.3%/83.9%, outperforming top-tier models and further boosting downstream table QA performance.

Keywords: merged-cell-aware table recognition · reinforced learning

1 Introduction

Tables are an efficient form of organizing structured data, widely found in various sources such as academic papers, financial reports, web pages, digital documents,

⋆Interns at Tongyi Lab, Alibaba Group, †Corresponding Author

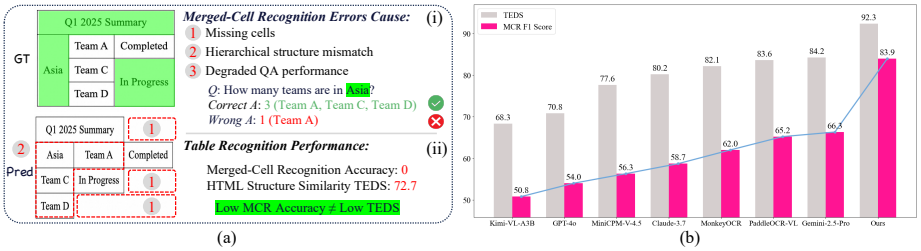

Fig. 1: (a) Incorrect merged-cell predictions can significantly degrade table structure recovery and QA accuracy, yet the evaluation metric TEDS, is largely insensitive to such errors. The green highlights in the GT indicate merged-cell regions. (b) Comparison of existing TR methods in terms of TEDS and merged-cell recognition F1 score.

and so on. Currently, the core research tasks in this domain include detection, recognition, extraction, and understanding [24, 29, 34, 51, 54, 58, 59].

This work focuses on Table Recognition (TR), aiming to parse table images into HTML sequences. Existing TR methods can be broadly categorized into two types: traditional Specialized Small Models (SSMs) [14, 24, 30], and Multimodal Large Language Models (MLLMs) [22, 31, 43]. Both SSMs and MLLMs exhibit strong performance in simple table scenarios, yet their ability to generalize to structurally complex tables, especially those with complex merged cells (colspan or rowspan > 1), is still limited [30, 31]. Empirical observations indicate that accurate prediction of merged-cell attributes is crucial for modeling structurally complex tables. For example, Fig. 1(a-i) shows that three incorrect merged cells introduce missing cells, disrupt hierarchical structures, and ultimately lead to erroneous answer in downstream table understanding tasks. Although several methods [7, 19, 23] recognize the importance of merged cells for TR and make corresponding improvements, they typically adopt TEDS [58] as the optimization objective and evaluation metric. However, as TEDS relies on the whole HTML tree edit distance, it is insensitive to fine-grained structural changes introduced by merged-cell errors. As shown in Fig. 1(a-ii), even in extreme cases where all merged-cell span attributes are predicted incorrectly (0% merged-cell accuracy), the TEDS score can still be as high as 72.7%. Motivated by this, to directly quantify merged-cell prediction quality, the Merged-Cell Recognition (MCR) F1 score is introduced as a dedicated subtask metric for TR. Fig. 1(b) shows that current methods achieve higher TEDS but lower merged-cell recognition accuracy. Even the advanced proprietary model Gemini-2.5-Pro reaches 84.2% TEDS, yet its MCR F1 is only 66.3%. This gap suggests that low-quality merged-cell recognition signals can be a key bottleneck preventing current TR methods from achieving better recognition performance.

We attribute the low merged-cell recognition quality to two main factors. Firstly, statistics in Fig. 3 reveal that tables with merged cells in existing open-source TR datasets [11, 24, 57] are relatively simple, with about 95% samples containing fewer than 10 merged cells, resulting in a highly imbalanced distri-

bution. Secondly, most current TR methods [16, 22, 31, 41] follow an end-to-end autoregressive paradigm, directly converting table images into HTML sequences via Supervised Fine-Tuning (SFT). However, in long-sequence HTML generation, such methods often overlook this fine-grained but structurally critical merged-cell information, leading to a drop in overall TR accuracy. To address the limitations of SFT, some approaches [19, 23] have explored directly using TEDS for global supervision, employing reinforcement learning to improve recognition quality. However, as discussed above, since TEDS is insensitive to merged-cell errors, these methods remain ineffective in fully resolving the problem.

To this end, we propose **Table-MCR2TR**, a comprehensive solution from both data and training perspectives, which leverages fine-grained yet critical MCR cues to better support the core table recognition objective. Firstly, to address the scarcity of tables exhibiting diverse and complex merged-cell structures, a tailored data pipeline is devised, incorporating both MLLM-based synthetic table images and web-crawled table images, followed by careful annotation and quality checking. The resulting dataset spans multiple domains, with a high proportion of large merged-cell-count tables, providing rich diversity for robust table recognition. Furthermore, to address the tendency of existing table recognition approaches [16, 22, 25, 31] to overlook merged cells, we propose Merged-Cell-Aware Table Recognition (MCATR), which explicitly injects merged-cell knowledge into MLLMs for parsing complex table structures more accurately. Beyond the core *TR* objective, the MCATR incorporates two cascaded tasks that explicitly model merged-cell semantics: (1) *Merged-Cell Recognition* learns the attributes of merged cells; and (2) *MCR2TR Alignment* leverages the learned MCR knowledge as a prior to facilitate subsequent table recognition. The training procedure first starts with supervised learning to inject alignment knowledge alongside TR and MCR capabilities into the model. Additionally, to better capture the complex merged-cell structure of tables, we go beyond token-level supervision and introduce sequence-level structural supervision. This is realized through merged-cell-aware collaborative reinforcement learning, which optimizes a holistic sequence-level objective guided by a tripartite reward comprising TEDS, MCR F1, and alignment score. Unlike prior approaches that treat MCR merely as an auxiliary training signal, MCATR is the first to both explicitly learn merged-cell information and effectively transfer it to final table recognition via the MCR2TR alignment task. We argue that this progressive recognition paradigm eases TR training and inference: instead of generating the entire HTML sequence in one shot, the model is first required to predict the critical MCR cues, which are then fed to the MLLM as contextual guidance for subsequent generation. This lowers the difficulty of table recognition and leads to more accurate HTML outputs. Experimental results show our Table-MCR2TR achieves SOTA performance on the average MCR F1 (83.9%) and TEDS score (92.3%), surpassing all open-source models and even the top-tier proprietary model Gemini-2.5-Pro, while also boosting downstream table QA performance with high-quality merged-cell and table recognition.

To summarize, our contributions are as follows:

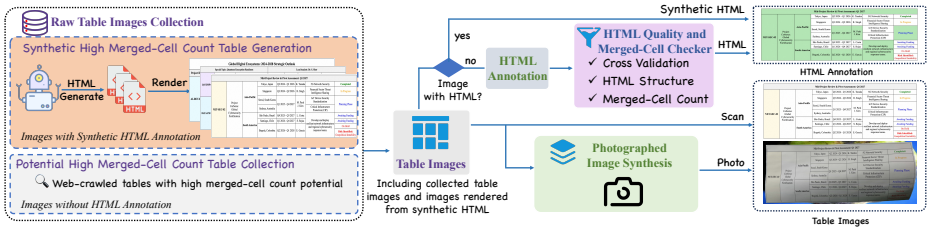

Fig. 2: Illustration of the proposed complex merged-cell table collection pipeline from MLLM-based synthetic table images and web-crawled table images.

- We propose a high-fidelity data construction pipeline that produces MLLM-generated and web-crawled tables with high merged-cell-counts, spanning multiple domains.
- We introduce Merged-Cell-Aware Table Recognition, which explicitly incorporates merged-cell information into MLLMs and adopts a progressive recognition strategy to more accurately parse complex table structures.
- By integrating this data construction pipeline with MCATR, the Table-MCR2TR achieves SOTA performance in terms of MCR and TR, and further enhances downstream table question answering performance.

2 Related Works

2.1 Existing Table Recognition Methods

Table recognition aims to convert visual tables into semi-structured text (*e.g.*, HTML or Markdown). Current mainstream methods can be categorized into two paradigms. The specialized small models paradigm [7, 18, 20, 21, 24, 28, 39, 51, 53, 58] typically adopts a divide-and-conquer strategy, decomposing the task into table structure recognition and table content recognition. These models feature smaller parameters, but suffer from limited capabilities in long sequence generation. When confronted with large-scale or structurally complex tables, the generated HTML sequences become excessively long, making it difficult for models to maintain global contextual consistency. The multimodal large language models [5,8,10,16,22,42,44,54,56,59], adopt an end-to-end approach, directly generating HTML or Markdown through prompts. Leveraging their massive parameter scale and extensive pre-training, MLLMs demonstrate powerful zero-shot or few-shot capabilities. However, beneath their impressive performance lies an implicit and unreliable understanding of table structures. For example, Dolphin [16] and MonkeyOCR [22] use SFT to directly map images to the HTML, lacking an intermediate step for explicit reasoning or validation of critical structural details like merged cells. To address this issue, this work introduces a merged-cell-aware table recognition solution that compels the model to develop a focused awareness of these critical merged cells and leverages the contextual understanding

capabilities of MLLMs to incorporate such information into subsequent table recognition, ultimately improving recognition accuracy.

2.2 Reinforcement Learning for MLLMs

Inspired by the success of reinforcement learning in pure language reasoning models, DeepSeek-R1 [17], recent research is actively exploring the transfer of RL paradigms to MLLMs to achieve performance beyond the limits of traditional supervised fine-tuning. In fundamental visual understanding tasks, VLM-RL [33] leverages Group Relative Policy Optimization (GRPO) [32] to directly optimize detection and localization rewards, thereby further enhancing the model's spatial reasoning capabilities and robustness in complex scenes. In the document intelligence domain, Infinity-Parser [40] designs layout-aware fine-grained rewards to achieve more precise document structure parsing. In MLLM for code generation, researchers such as Table2LaTeX-RL [23] and MSRL [6] have also improved the accuracy of generated LaTeX or Python code by introducing visual consistency rewards. However, existing works primarily focus on designing rewards for the core task, overlooking how intermediate steps affect the final outcome. Building upon this foundation, we go a further step by exploring a merged-cell-aware collaborative reward mechanism, which not only optimizes the reward for the core task but also ensures the quality of the intermediate MCR process.

3 Method

In this section, our Table-MCR2TR is presented, comprising a complex merged-cell table collection pipeline and a novel training paradigm (MCATR) that explicitly incorporates merged-cell generation for superior table recognition.

3.1 Complex Merged-Cell Table Data Collection

To generate large-scale table data with complex merged-cell structures and reliable annotations, we design a two-branch pipeline (Fig. 2) integrating MLLM-synthesized and web-crawled tables, followed by HTML annotation, quality checking, and photorealistic image synthesis. This unified process produces high-quality data across scanned and photo domains, effectively addressing the scarcity of complex merged-cell tables while ensuring high-fidelity annotations.

Raw Table Images Collection. First, synthetic high merged-cell count table generation uses MLLMs (*e.g.*, Gemini-2.5-Pro) with prompts explicitly controlling merged-cell count, rowspan, and colspan to generate complex HTML tables. The generated HTML is then rendered into scanned-style images to align structure and appearance. Second, potential high merged-cell count table collection gathers web-crawled table images without HTML annotation, ensuring broad coverage across domains such as financial statements, accounting reports, insurance forms, and customs declarations. Unlike synthetic tables, which are labeled

during generation, web-crawled tables require automated HTML annotation via advanced MLLMs (*e.g.*, Qwen-VL-max, PaddleOCR-VL, MinerU, Gemini, *etc.*).

HTML Quality and Merged-Cell Checker. Following the preliminary HTML annotations of the web-crawled tables from the previous stage, this module focuses on improving annotation quality and identifying HTML tables with a large number of merged cells.

1) Cross-Model Consistency Checker. For each web-crawled table image, we run multiple MLLMs to obtain a set of predicted HTML sequences $\{H_i\}_{i=1}^{K}$, where K denotes the number of recognition models. We then compute the pairwise TEDS scores between all sequence pairs, and derive the mean: $\text{TEDS}_{avg} = \frac{2\sum_{1\le i<j\le K}\text{TEDS}(H_i,H_j)}{K(K-1)}$. Since a low mean TEDS_{avg} suggests ambiguous and degraded table structure, we discard samples with $\text{TEDS}_{avg} < \mu_{TEDS}$, retaining only high-confidence annotations.

2) HTML Structure Checker. This component evaluates the structural coherence of the predicted HTML table. First, the HTML sequence is parsed into a 2D logical grid, where each occupied cell is marked as 1, and unoccupied cells are marked as 0. If the proportion of unoccupied cells exceeds τ_{frag} of the total grid size, the HTML is considered fragmented and is filtered out.

3) Merged-Cell Count Checker. After the consistency and structure checks, we compute the number of merged cells per table and filter accordingly: tables with more than 10 merged cells are kept, while the rest are sampled with probability. This ensures a high proportion of large merged-cell-count tables in the dataset, thereby strengthening the model's ability to handle complex table structures.

Photographed Image Synthesis. To extend scanned table data into complex and realistic photo scenarios, photorealistic images are synthesized from clean scans by inverting document unwarping methods [15, 38]. Whereas unwarping flattens photographed documents, the reverse process warps clean scans into photo-like tables. UVDoc [38] provides real-world geometries P and UV maps U. Given a scanned table image I, we randomly sample P and generate a textured warped image via $\mathcal{M}$: $I_p = \mathcal{M}(I, P, U)$, where $\mathcal{M}$ performs UV warping with multiplicative blending. We then composite I_p onto a random background B with color matching using $\mathcal{C}$: $\hat{I} = \mathcal{C}(I_p, B, U)$. Compared to rendering-based approaches [47, 48], this method yields more realistic images by leveraging UV maps and geometries captured by depth camera measurements and manual annotation.

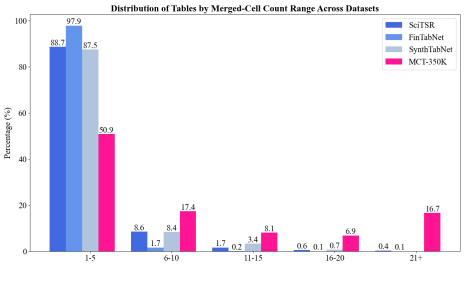

Fig. 3: Distribution of tables by merged-cell count range across existing datasets and our MCT-350K.

MCT-350K Dataset Analysis. Through the above pipeline, we construct a large-scale table recognition dataset with abundant merged cells, named MCT-

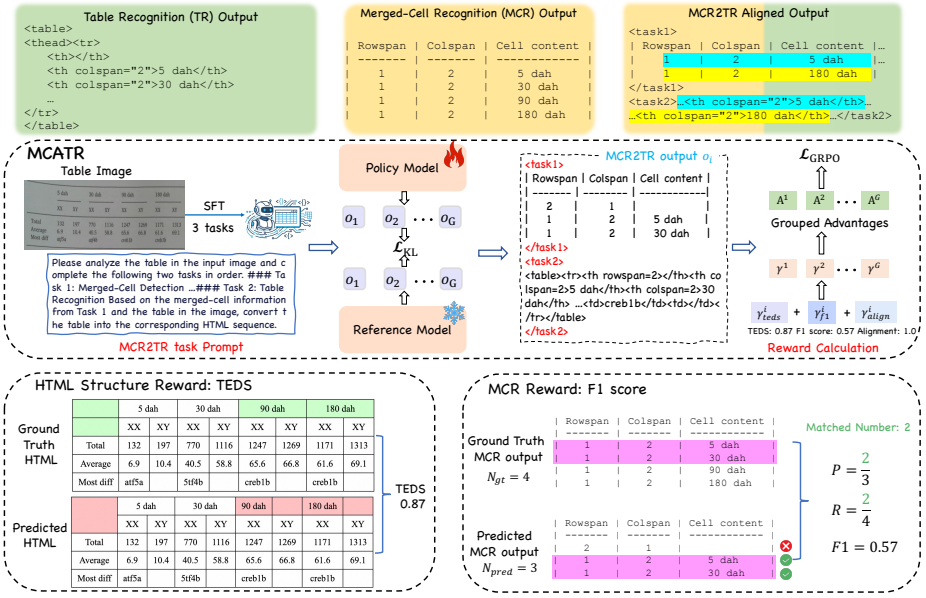

Fig. 4: The overall framework of MCATR. (Top) Three tasks are defined: Table Recognition, Merged-Cell Recognition, and MCR2TR Alignment. (Middle) SFT initializes the model, followed by merged-cell-aware collaborative reinforcement learning. (Bottom) Rewards combine TEDS, MCR F1, and alignment scores.

350K, and the distribution of merged-cell counts is shown in Fig. 3. We intentionally increase the proportion of tables with more than 10 merged cells, especially for those with counts exceeding 20, which account for 16.7%. The MCT-350K tables average 12 rows and 6 columns with 1623×1351 resolution. In addition, the dataset exhibits rich linguistic diversity, covering Chinese, English, and multilingual tables, and spans both photographed and scanned domains.

3.2 MCATR Framework

Building upon the generated MCT-350K and existing open-source table data, we introduce a novel training framework for Merged-Cell-Aware Table Recognition **MCATR**. As illustrated in Fig. 4, the process begins with supervised fine-tuning and culminates in a merged-cell-aware collaborative reinforcement learning (MCRL) stage. Unlike previous approaches [5,24,25,58], MCATR reformulates TR by introducing merged-cell recognition as an explicit auxiliary task to encode structural priors. Conditioning HTML generation on these merged-cell priors substantially improves the model's ability to capture complex table layouts.

Supervised Fine-Tuning. As shown in the upper part of Fig. 4, in the first training stage, our MCATR is structured around three complementary tasks. The core objective, Table Recognition (TR), transcribes a table image into its corresponding HTML sequence. Then, to enhance the awareness of merged-cell

structures in long HTML sequences, Merged-Cell Recognition (MCR) is introduced as an auxiliary task. MCR jointly predicts row and column spans together with the textual content of each merged cell, enabling a more accurate and robust structural representation in table recognition. Furthermore, to bridge the gap between the standalone MCR task and the core TR objective, and to ensure that merged-cell information genuinely benefits TR, the MCR2TR alignment task is proposed that enforces cross-task consistency via a progressive recognition process: first recognizing merged cells, and then generating HTML conditioned on the MCR output. With this SFT procedure, the model is encouraged to explicitly attend to fine-grained merged-cell information and is also equipped with format knowledge needed for the alignment task. By predicting merged-cell cues before table recognition and using them as context for the TR model, we effectively reduce the difficulty of direct TR generation.

Merged-cell-aware Collaborative Reinforcement Learning. Building on supervised learning, our model effectively learns table recognition, merged-cell recognition, and their alignment patterns, capturing both tag-level formats and domain-specific structural knowledge. However, its token-level auto-regressive objective inherently focuses on local predictions, which limits the ability to represent the nested and hierarchical syntax characteristic of HTML tables, especially in complex merged-cell layouts. To further enhance global structural fidelity, a GRPO-based merged-cell-aware collaborative reinforcement fine-tuning is introduced, which conducts sequence-level optimization over complete table hypotheses. This enables the model to explore diverse layout configurations and converge toward solutions that are both globally optimal and structurally coherent. Moreover, to effectively guide M-CRL, a tripartite set of task-specific reward function is designed: TEDS, MCR F1, and alignment score.

To evaluate the overall structural quality of a predicted HTML structure, TEDS is used. Specifically, TEDS first computes the minimum edit distance $\mathrm{EditDist}(\cdot,\cdot)$ between two HTML trees by applying a series of operations (i.e., node insertion, deletion, and substitution), and then normalizes this distance to yield a similarity score in $[0,1]$. Given a predicted HTML sequence H_{pred} and its ground-truth counterpart H_{gt}, the TEDS reward is defined:

$$\mathrm{TEDS}(H_{\text{pred}}, H_{\text{gt}}) = 1 - \frac{\mathrm{EditDist}(H_{\text{pred}}, H_{\text{gt}})}{\max(|H_{\text{pred}}|, |H_{\text{gt}}|)}, \tag{1}$$

where $|\cdot|$ represents the number of nodes in the HTML tree.

To evaluate the agreement between predicted and ground-truth merged-cell information, the MCR F1 score is defined as reward, with its implementation outlined in Algorithm 1. Specifically, the predicted merged cells are post-processed into a list: $M_{\text{pred}} = \{(r_1, c_1, t_1), (r_2, c_2, t_2), \ldots, (r_i, c_i, t_i)\}$, where r_i, c_i, and t_i denote the rowspan, colspan, and textual content of the i-th predicted merged-cell, respectively. Similarly, the ground-truth merged-cell annotations are formatted into an equivalent list M_{gt}. Let $Match_{num}$ denote the count of $p \in M_{pred}$ such that $(p[0], p[1], p[2])$ matches an element in M_{gt} in rowspan, colspan, and content similarity. Finally, $Match_{num}$ is converted to the MCR F1 score reward using precision and recall definition.

Algorithm 1 Pseudo Code for MCR F1 Score Calculation

▷ *Input: List of predicted merged cells:* M_{pred}*, List of ground-truth merged cells:* M_{gt}*, Content similarity threshold:* τ_{sim}
▷ *Output: MCR F1 Score*
$Match_{num}, Match_{set} \leftarrow 0, \{\}$
for each p in M_{pred} **do**
    $Match_{item}, max_{sim} \leftarrow \text{null}, -1$
    for each g in M_{gt} & g not in $Match_{set}$ **do**
        $sim \leftarrow \text{ANLS}(p[2], g[2])$ ▷ cell content similarity
        if $p[0] = g[0]$ & $p[1] = g[1]$ & $sim > \tau_{sim}$ **then** ▷ merged cell matching
            if $sim > max_{sim}$ **then**
                $max_{sim} \leftarrow sim, \quad Match_{item} \leftarrow g$
    if $Match_{item}$ is not null **then** ▷ merged cell matched
        $Match_{num} \leftarrow Match_{num} + 1, \quad$ Add $Match_{item}$ to $Match_{set}$
$P \leftarrow Match_{num}/\text{len}(M_{pred}), \quad R \leftarrow Match_{num}/\text{len}(M_{gt})$
$F1 \leftarrow \frac{2 \times P \times R}{P+R}$
return $F1$

To facilitate the MCR2TR alignment objective, we define an alignment reward that enforces consistency between the merged-cell attributes in the final HTML output and those predicted merged cells in the first stage. Specifically, we extract merged-cell information from the outputs of the MCR and TR stages, denoted as M_{gt} and M_{pred} respectively, and compute the alignment reward.

Reward Calculation. Let r_{teds}, r_{f1}, and r_{align} denote the TEDS, MCR F1 score, and alignment reward, respectively. The task-specific rewards are: $r_{\text{TR}} = r_{\text{teds}}$, $r_{\text{MCR}} = r_{\text{f1}}$, $r_{\text{MCR2TR}} = r_{\text{teds}} + r_{\text{f1}} + r_{\text{align}}$. This formulation follows directly from the evaluation objectives: the TR task emphasizes table-level structural similarity, the MCR task targets merged-cell accuracy, and the joint alignment task integrates both structure and merged-cell rewards, with r_{align} ensuring MCR outputs can be effectively utilized by the subsequent TR stage.

4 Experiments

4.1 Experimental Setup

Training Datasets. In addition to our constructed MCT-350K dataset, which features a large proportion of complex tables with numerous merged cells and diverse structural patterns, several open-source datasets (SciTSR [11], PubTabNet [58], FinTabNet [57], TabRectSet [49], and SynthTabNet [24]) are also used. During the SFT stage, the three TR, MCR, and MCR2TR alignment tasks are trained on these million tables each. In the reinforcement fine-tuning stage, 7K distinct table images are selected and shared across all three tasks, yielding a total of 21K training samples.

Implementation Details. In the HTML quality check stage, the thresholds are set to $\tau_{frag} = 0.05$, $\mu_{\text{TEDS}} = 0.7$. For the MCR F1 score, the similarity

Table 1: Evaluation results across all benchmarks. F1 refers to the MCR F1 score. For the Scan and Photo of CC-OCR, results are reported as MCR F1 score/TEDS (xx/yy) due to space constraints. **Bold** and underlined values indicate the best and second-best results, respectively.

Methods	SciTSR		PubTabNet		FinTabNet		SynthTabNet		TabRectSet		CC-OCR		OmniDocbench		AVG	
	F1	TEDS	F1	TEDS	F1	TEDS	F1	TEDS	F1	TEDS	Scan	Photo	F1	TEDS	F1	TEDS
Specialized Small Models																
RapidTable [30]	88.3	87.7	**81.5**	86.8	67.0	77.1	16.6	74.2	60.7	63.0	56.3/66.4	37.3/39.0	74.1	83.1	60.2	72.2
PP-StructureV3 [14]	87.5	82.7	68.9	73.1	42.9	58.6	13.1	59.3	62.2	55.8	49.7/65.0	35.7/33.6	74.0	81.3	54.3	63.7
General MLLMs																
Qwen2.5-VL-32B [5]	80.2	92.9	51.2	78.6	49.4	73.6	0.5	61.7	58.1	73.1	53.4/84.8	53.1/79.4	68.5	86.6	51.8	78.8
Qwen3-VL-32B [4]	89.3	94.1	58.5	84.5	56.9	84.8	5.0	70.2	61.5	73.6	58.0/86.5	52.0/78.9	80.0	90.7	57.7	83.0
InternVL3.5-8B [43]	85.5	93.8	56.7	83.9	49.3	88.1	2.9	77.5	56.2	71.4	59.4/82.1	46.3/68.1	67.9	85.0	53.0	81.2
InternVL3.5-38B [43]	88.2	94.2	62.7	87.5	71.8	92.8	6.4	81.0	58.9	75.8	56.7/84.8	43.6/71.6	71.3	86.2	57.5	84.2
DeepSeek-VL2 [46]	59.0	82.0	50.5	81.0	45.5	79.4	0.3	53.7	43.8	61.2	47.3/67.1	42.0/53.4	52.7	70.8	42.6	68.6
MiniCPM-V-4.5 [52]	84.9	90.5	61.2	79.4	65.4	86.8	0.4	61.0	56.1	70.3	59.6/79.2	49.7/69.1	73.1	84.7	56.3	77.6
Gemma-3-27B [36]	76.5	89.4	51.8	81.6	49.7	78.8	0.2	58.2	55.0	68.7	51.0/64.7	48.0/53.8	62.6	75.5	49.4	71.3
Kimi-VL-A3B [37]	80.9	88.0	51.4	72.3	51.7	69.3	0.4	52.3	53.0	64.9	56.8/67.9	48.2/54.6	63.4	77.1	50.8	68.3
Specialized MLLMs																
dots.ocr [31]	83.2	86.9	56.8	88.3	76.4	71.1	3.7	62.8	61.7	62.8	71.3/81.3	56.0/64.3	78.7	81.0	61.0	74.8
MinerU2-VLM [41]	89.5	93.0	68.7	90.3	60.7	84.4	73.1	94.4	62.9	73.9	73.4/75.1	54.3/46.2	83.0	89.8	70.7	80.9
MonkeyOCR [22]	83.1	90.6	68.5	87.9	54.0	82.6	59.6	93.1	56.7	71.6	49.6/76.3	48.8/70.4	75.9	84.3	62.0	82.1
PaddleOCR-VL [13]	90.6	92.7	68.1	87.1	59.1	83.5	28.1	79.7	68.9	77.4	62.9/81.3	59.2/74.7	**84.3**	**92.0**	65.2	83.6
DeepSeek-OCR [45]	88.0	89.9	72.3	91.1	92.4	96.5	4.8	72.6	59.7	75.6	69.3/80.6	49.4/56.1	76.3	85.0	64.0	80.9
Closed MLLMs																
GPT-4o [1]	81.9	88.9	59.3	73.2	56.8	79.8	2.8	50.1	57.0	70.8	57.7/65.5	47.3/57.6	69.1	80.2	54.0	70.8
Gemini-1.5-Pro [35]	87.4	92.4	65.0	84.7	59.3	88.3	1.7	67.1	59.9	78.5	59.1/77.0	50.0/67.9	74.2	86.4	57.1	80.3
Gemini-2.5-Pro [12]	91.4	92.9	72.9	90.9	79.7	92.8	8.7	54.4	72.2	80.6	69.6/89.4	54.6/ 82.2	81.5	91.0	66.3	84.2
Claude-3.7-Sonnet [3]	92.2	94.3	59.8	74.0	70.7	82.6	0.2	68.8	61.7	75.2	57.4/85.2	49.3/74.6	78.0	86.9	58.7	80.2
Table-MCR2TR	**95.5**	**97.7**	80.6	**93.7**	**95.7**	**97.0**	**87.6**	**98.7**	**79.0**	**86.5**	**76.6/90.5**	**71.2/83.0**	83.6	91.0	**83.9**	**92.3**

threshold $\tau_{sim} = 0.5$. We perform full-parameter tuning across all stages, using Qwen2.5-VL-3B as the base model. During the 3-task joint SFT phase, the maximum image token length is set to 2048, and the total context length (including image and text) is capped at 10800. Training is conducted on 2 nodes (16×A100 GPUs), with a batch size of 8 and gradient accumulation over 2 steps. For the reinforcement learning stage, we adopt the MS-Swift framework [55], with training also conducted on 2 nodes. Each sample is rolled out 8 times (num_generations = 8), with a batch size of 2 and gradient accumulation over 4 steps. The maximum output length is set to 6144 tokens. We use the temperature = 1.0, top_k = 50, and top_p = 0.9 to encourage output diversity. The KL divergence loss weight β between the policy and reference models is set to 0.04.

Evaluation. To comprehensively evaluate table recognition, we conduct experiments on multiple benchmarks. Given the scale of SynthTabNet (nearly 60K test samples), we randomly sample 10K for evaluation to reduce cost. For SciTSR (3K), PubTabNet (9K), FinTabNet (10K), and TabRectSet (1.2K), we evaluate on the full test sets. We further report results on OmniDocBench table sets [26] (428 images) and the more challenging CC-OCR table sets [50] (150 scanned and 150 photographed images). Except for the closed-source models, all other models are evaluated uniformly on the aforementioned test sets. For closed-source models, we evaluate on 500 randomly sampled instances from each of SciTSR, PubTabNet, FinTabNet, and SynthTabNet, while using the full test sets for the remaining benchmarks. In all MLLM evaluations, we set the temperature to 0 to ensure output stability, and cap the maximum output length at 8192 tokens. Performance is evaluated using both TEDS and MCR F1 score.

4.2 Main Results

In Tab. 1, the Table-MCR2TR is compared with a wide range of mainstream methods: traditional SSMs, general and document-specific MLLMs, and leading proprietary models, with the last two columns (AVG) reporting the mean of the MCR F1 score and TEDS.

Comparison with Specialized Small Models. Table-MCR2TR significantly outperforms existing SSMs, including RapidTable [30] and PP-StructureV3 [14], across multiple benchmarks (*e.g.*, SciTSR, FinTabNet, OmniDocBench, and the more challenging CC-OCR) on MCR F1 score and TEDS. On PubTabNet, it lags behind RapidTable by only 0.9% on the MCR F1 score; this gap is attributed to PubTabNet's low image quality that adversely affects merged-cell recognition. Nevertheless, thanks to the MLLM's superior text recognition capability, Table-MCR2TR achieves the best TEDS of 93.7%.

Comparison with General MLLMs. We compare Table-MCR2TR against multiple mainstream general-purpose MLLMs [4,5,36,37,43,46,52]. As shown in Tab. 1, all general MLLMs underperform Table-MCR2TR across every benchmark. Even the strong InternVL3.5-VL-38B attains an average TEDS of 84.2%, lagging behind the proposed method's 92.3% by 8.1%. Interestingly, although it attains a high TEDS of 81.0% on SynthTabNet, its merged-cell recognition accuracy is only 6.4%, indicating that almost all merged cells are predicted incorrectly. This phenomenon arises from the fact that nearly all tables in the SynthTabNet test set contain merged cells, while these methods tend to disregard such fine-grained structural information, resulting in low MCR F1 scores. In addition, the TEDS of the weakest MLLM Kimi-VL-A3B falls behind that of our model by up to 24%, while the gap in MCR F1 score reaches 33%. These large performance gaps further demonstrate the effectiveness of our approach, which enhances overall table recognition quality through explicit merged-cell-aware modeling.

Comparison with Specialized MLLMs. As shown in Tab. 1, document-specialized MLLMs [13, 22, 31, 41, 45] achieve an average TEDS above 80% due to training on large-scale table data. However, their merged-cell recognition quality remains largely unimproved, which constrains their overall TR performance. Although these models are trained with supervision on large-scale table recognition data, they still pay insufficient attention to fine-grained merged-cell cues. As a result, they perform poorly on the complex CC-OCR set in terms of both TEDS and F1 score, with a substantial gap compared to our model. This also suggests that, for MLLMs, relying solely on SFT to directly supervise the entire HTML sequence remains challenging. In contrast, our method explicitly treats accurate merged-cell information as crucial for TR, and introduces MCR together with the alignment task so that the model does not have to produce the HTML sequence in one shot. Instead, it follows a chain-of-thought style: it first focuses on fine-grained MCR and then uses MCR outputs as contextual priors, enabling MLLMs to leverage their own reasoning capability to facilitate final table recognition.

Table 2: The Effectiveness of the proposed data constructed pipeline and learning strategy. MCT-350K and MCATR denote the proposed complex merged-cell table dataset and framework, respectively.

#	T-350K	MCATR	Scan		Photo		AVG	
			F1	TEDS	F1	TEDS	F1	TEDS
1	-	-	44.7	77.9	48.1	75.4	46.4	76.7
2	✓	-	68.7	87.1	59.2	79.8	64.0	83.5
3	✓	✓	76.6	90.5	71.2	83.0	**73.9**	**86.8**

Table 3: Ablation results of rewards in MCR2TR aligned task.

#	MCR2TR Rewards			Scan		Photo		AVG	
	Align	TEDS	F1	F1	TEDS	F1	TEDS	F1	TEDS
1	3-task joint SFT			73.3	88.6	62.5	80.9	67.9	84.8
2	✓	✓	-	74.9	89.7	66.4	82.2	70.7	86.0
3	✓	-	✓	75.5	89.5	69.4	81.9	72.5	85.7
4	-	✓	✓	76.0	90.1	70.4	82.6	73.2	86.4
5	✓	✓	✓	76.6	90.5	71.2	83.0	**73.9**	**86.8**

Comparison with Closed MLLMs. Our Table-MCR2TR also outperforms closed models (GPT-4o [1], Gemini-1.5-Pro [35], and Claude-3.7-Sonnet [3]) across all test sets, with an average TEDS gain of at least 12.0%. Even compared with the top-tier proprietary model Gemini-2.5-Pro [12], Table-MCR2TR achieves superior results in most scenarios, with average improvements of 8.1% in TEDS and 17.6% in MCR F1 score across diverse benchmarks, including both relatively simple academic datasets and the more challenging CC-OCR benchmark. It is worth noting that these closed-source models typically have much larger parameter counts and are trained on substantially more data, further highlighting the advantage conferred by our constructed complex table dataset and the proposed MCATR framework.

4.3 Ablation Study

Effectiveness of Data Pipeline and MCATR. The results in Tab. 2 show that the base model Qwen2.5-VL-3B (#1) performs poorly on the Photo and Scan subsets of the challenging CC-OCR benchmark. Incorporating our constructed complex merged-cell table dataset MCT-350K along with open-source table data for TR-only SFT training (#2) improves overall table recognition performance; however, the MCR F1 score remains low, indicating that TR-only training is insufficient to capture fine-grained merged-cell structural details during HTML generation. Finally, incorporating the proposed MCATR framework (#3) explicitly models merged-cell recognition, yielding further gains in both MCR F1 score and overall TEDS performance.

Rewards in MCR2TR Alignment Task. As shown in Tab. 3, training with 3-task joint SFT (#1) yields averages of 67.9% MCR F1 score and 84.8% TEDS.

Table 4: The importance of merged-cell recognition quality for table recognition.

MCR information	Scan	Photo	OmniDocbench
Our pred MCR as input	90.5	83.0	91.0
GT MCR as input	93.9 (+3.4)	87.6 (+4.6)	93.6 (+2.6)

Table 5: Ablation results of scaling to other models.

3-task SFT	M-CRL	Scan		Photo		AVG	
		F1	TEDS	F1	TEDS	F1	TEDS
InternVL3-1B		48.5	59.5	46.3	47.9	47.4	53.7
✓	-	65.2	82.5	51.3	66.8	58.3	74.7
✓	✓	68.5	85.7	60.3	74.6	**64.4**	**80.2**
Qwen3-VL-2B		48.0	67.6	46.3	63.4	47.2	65.5
✓	-	70.2	87.1	61.3	79.6	65.8	83.4
✓	✓	74.6	90.3	71.5	85.4	**73.1**	**87.9**
Qwen2.5-VL-3B		44.7	77.9	48.1	75.4	46.4	76.7
✓	-	73.3	88.6	62.5	80.9	67.9	84.8
✓	✓	76.6	90.5	71.2	83.0	**73.9**	**86.8**

Adding either TEDS (#2) or MCR F1 score (#3) alongside alignment reward improves the MCR F1 score by 2.8% and 4.6%, respectively, confirming that explicit merged-cell supervision is more effective than TEDS for better MCR. Removing alignment reward (#4) leads to a drop in F1 to 73.2%, demonstrating its necessity. Combining all three rewards (#5) delivers the best results (73.9% F1, 86.8% TEDS), highlighting their synergistic effects in preserving structural correctness and merged-cell recognition accuracy. We argue that incorporating the MCR F1 score in the MCR2TR alignment task provides a clearer learning signal for reinforcement learning. This is because MCR F1 score mainly relies on discrete matching logic and tends to be relatively sparse, whereas TEDS is comparatively dense and often yields near-zero reward variance during rollout, resulting in weak optimization signals. In contrast, MCR F1 score more strongly rewards rollout samples that correctly identify merged cells, thereby enlarging the advantage gaps among candidate samples and offering a more discriminative objective. This complementary signal also mitigates TEDS's insensitivity to merged-cell errors and better guides the policy model toward improvement.

Impact of MCR Quality on TR Performance. The MCR2TR alignment task requires the model to first generate MCR signals and use them as explicit conditions for subsequent TR. This is not a naive concatenation of MCR and TR outputs; rather, it introduces an intrinsic dependency between the two tasks. As discussed earlier, existing models often struggle to capture fine-grained yet important merged-cell cues, which can substantially affect table recognition. Our alignment task therefore extracts such critical signals upfront and injects them as contextual priors, reducing the difficulty of the subsequent TR. Consequently, more accurate MCR predictions consistently translate into better performance.

Table 6: Downstream table QA performance. **Left**: our first parsing tables then QA pipeline (Ours → HTML → LLM) vs. direct MLLM-based table QA methods. **Right**: different TR methods evaluated with the same QA LLM. Q2/Q3/DS-OCR/PubTab denote Qwen2, Qwen3, DeepSeek-OCR, and PubHealthTab, respectively.

Method	WTQ	TabFact	PubTab
GPT-4o	64.5	79.0	65.5
Gemini-2.5	76.5	78.0	69.0
QVQ-72B	72.0	77.0	67.0
Q3-VL-235B	70.0	80.0	68.1
Ours+Q3-8B	76.8	89.2	74.1
Ours+Q3-32B	**77.6**	**89.8**	**75.8**

Method	WTQ	TabFact	PubTab
QA Model Qwen3-32B			
GPT-4o	67.7	86.6	72.6
MinerU2	75.2	87.4	73.0
DS-OCR	66.9	79.6	71.4
Ours	**77.6**	**89.8**	**75.8**

Tab. 4 shows that if our model can correctly predict the ground-truth merged-cell information (100% MCR F1 score), its performance further improves, with substantial gains on both CC-OCR (+4.0%) and OmniDocBench (+2.6%).

Scale to Other Base Models. We also evaluate the generalization ability of the proposed data collection pipeline and MCATR framework on models with 1B, 2B, and 3B parameters. As shown in Tab. 5, the base models perform poorly; for example, InternVL3-1B [60] records only 47.4% MCR F1 score and 53.7% TEDS. Applying supervised fine-tuning with our MCT-350K substantially boosts performance across all model sizes, yielding gains of up to 21% in TEDS. Further incorporating M-CRL provides additional improvements, achieving 64.4/80.2 (1B), 73.1/87.9 (2B), and 73.9/86.8 (3B) in F1/TEDS. These consistent gains across varying capacities underscore the scalability and effectiveness of our table construction pipeline and the MCATR framework.

4.4 Application to Downstream Table Understanding Tasks

Beyond table recognition, our Table-MCR2TR enables a two-stage multimodal table question answering (QA) pipeline: it first parses the table image into an HTML sequence and then feeds this structured representation along with the question to a language model for answer generation. We evaluate on three multimodal table QA benchmarks: WTQ [27], TabFact [9], and PubHealthTab [2]. As shown in Tab. 6 (left), our two-stage pipeline (last two rows) achieves the best performance across all benchmarks despite using a much smaller language model (Qwen3-32B), which underscores the high-fidelity of our table structure parsing due to accurate MCR quality. Moreover, the right table isolates the impact of the parsing methods by pairing different table parsers with the same QA model. Under this controlled setting, Table-MCR2TR consistently yields the highest accuracy, outperforming GPT-4o, MinerU2-VLM, and DeepSeek-OCR. The above results demonstrate the superiority of our approach. By accurately recognizing merged cells, the model can better recover both the structure and content of table images in the subsequent TR stage, thereby minimizing spatial misalignment and structural mismatches. Consequently, the high-fidelity HTML sequences produced by our method provide reliable structured inputs for downstream table understanding, leading to more accurate answers.

4.5 Discussion on Inference Efficiency and General Capability

For inference efficiency, all experiments are deployed on 4×A100 GPUs. Our Table-MCR2TR retains around 90% of the base model throughput without explicit MCR prediction (3.7 *vs*. 4.1 tables/s), indicating that MCR sequence generation introduces only a marginal inference slowdown. Regarding the influence on general MLLM ability, we observe that after 3-task SFT and M-CRL training, the model can no longer reliably follow general-purpose instructions and only responds as expected to the three prompts used during training, resulting in an almost complete loss of general capability. This degradation is likely a consequence of intensive TR-focused fine-tuning on a large domain-specific dataset, which drives strong task specialization. Such a trade-off between generality and specialization has also been seen in other document-specialized MLLMs.

5 Conclusion

We present Table-MCR2TR, a reinforced MLLM framework that tackles both the scarcity of complex table data with high merged-cell counts and the inaccurate merged-cell recognition during HTML generation. It features a high-fidelity pipeline that constructs the MCT-350K dataset using a combination of MLLM-synthesized and web-crawled tables, and a merged-cell-aware optimization strategy, MCATR, which jointly optimizes three complementary tasks: Table Recognition, Merged-Cell Recognition, and MCR2TR alignment. Through large-scale supervised learning and merged-cell-aware collaborative reinforcement learning, Table-MCR2TR can both explicitly perform MCR and leverage the MLLM's reasoning ability to use this information for final table recognition. Our Table-MCR2TR achieves SOTA performance on multiple TR benchmarks and demonstrates strong scalability. Moreover, thanks to high-quality MCR, the generated HTML sequences also yield positive gains for downstream table QA tasks.

Acknowledgements. This work is supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM103), and the National Nature Science Foundation of China (62425114, 62121002, U23B2028). We thank the Alibaba Research Intern Program for supporting GPU resources and closed-source model APIs access. We also acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC and the USTC supercomputing center for providing computational resources for this project.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

2. Akhtar, M., Cocarascu, O., Simperl, E.: PubHealthTab: A public health table-based dataset for evidence-based fact checking. In: Findings of the Association for Computational Linguistics: NAACL 2022. pp. 1–16. Association for Computational Linguistics, Seattle, United States (Jul 2022)
3. Anthropic: The claude 3 model family: Opus, sonnet, haiku (2025)
4. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
5. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
6. Chen, L., Zhao, X., Zeng, Z., Huang, J., Zheng, L., Zhong, Y., Ma, L.: Breaking the sft plateau: Multimodal structured reinforcement learning for chart-to-code generation. arXiv preprint arXiv:2508.13587 (2025)
7. Chen, L., Huang, C., Zheng, X., Lin, J., Huang, X.J.: Tablevlm: multi-modal pre-training for table structure recognition. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2437–2449 (2023)
8. Chen, Q., Zhang, X., Guo, L., Chen, F., Zhang, C.: Dianjin-ocr-r1: Enhancing ocr capabilities via a reasoning-and-tool interleaved vision-language model. arXiv preprint arXiv:2508.13238 (2025)
9. Chen, W., Wang, H., Chen, J., Zhang, Y., Wang, H., Li, S., Zhou, X., Wang, W.Y.: Tabfact: A large-scale dataset for table-based fact verification. In: International Conference on Learning Representations (2019)
10. Chen, X., Li, S., Zhu, X., Chen, Y., Yang, F., Fang, C., Qu, L., Xu, X., Wei, H., Wu, M.: Logics-parsing technical report. arXiv preprint arXiv:2509.19760 (2025)
11. Chi, Z., Huang, H., Xu, H.D., Yu, H., Yin, W., Mao, X.L.: Complicated table structure recognition. arXiv preprint arXiv:1908.04729 (2019)
12. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
13. Cui, C., Sun, T., Liang, S., Gao, T., Zhang, Z., Liu, J., Wang, X., Zhou, C., Liu, H., Lin, M., et al.: Paddleocr-vl: Boosting multilingual document parsing via a 0.9 b ultra-compact vision-language model. arXiv preprint arXiv:2510.14528 (2025)
14. Cui, C., Sun, T., Lin, M., Gao, T., Zhang, Y., Liu, J., Wang, X., Zhang, Z., Zhou, C., Liu, H., et al.: Paddleocr 3.0 technical report. arXiv preprint arXiv:2507.05595 (2025)
15. Das, S., Ma, K., Shu, Z., Samaras, D., Shilkrot, R.: Dewarpnet: Single-image document unwarping with stacked 3d and 2d regression networks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 131–140 (2019)
16. Feng, H., Wei, S., Fei, X., Shi, W., Han, Y., Liao, L., Lu, J., Wu, B., Liu, Q., Lin, C., et al.: Dolphin: Document image parsing via heterogeneous anchor prompting. arXiv preprint arXiv:2505.14059 (2025)
17. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
18. Jung, G., Tang, Q., Lee, H., Jung, H.: Mixloss: Table structure recognition in industrial documents via collaborative learning. In: 2025 17th International Conference on Human System Interaction (HSI). pp. 1–6. IEEE (2025)

19. Kang, X., Wu, S., Wang, Z., Liu, Y., Jin, X., Huang, K., Wang, W., Yue, Y., Huang, X., Wang, Q.: Can grpo boost complex multimodal table understanding? In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 12642–12655 (2025)
20. Lee, E., Jo, Y., Lee, S., Kim, S.H., Cho, N.I.: Single-stage table structure recognition approach via efficient sequence modeling. Pattern Recognition Letters (2025)
21. Li, C., Guo, R., Zhou, J., An, M., Du, Y., Zhu, L., Liu, Y., Hu, X., Yu, D.: Pp-structurev2: A stronger document analysis system. arXiv preprint arXiv:2210.05391 (2022)
22. Li, Z., Liu, Y., Liu, Q., Ma, Z., Zhang, Z., Zhang, S., Guo, Z., Zhang, J., Wang, X., Bai, X.: Monkeyocr: Document parsing with a structure-recognition-relation triplet paradigm. arXiv preprint arXiv:2506.05218 (2025)
23. Ling, J., Qi, Y., Huang, T., Zhou, S., Huang, Y., Yang, J., Song, Z., Zhou, Y., Yang, Y., Shen, H.T., et al.: Table2latex-rl: High-fidelity latex code generation from table images via reinforced multimodal language models. arXiv preprint arXiv:2509.17589 (2025)
24. Nassar, A., Livathinos, N., Lysak, M., Staar, P.: Tableformer: Table structure understanding with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4614–4623 (2022)
25. Niu, J., Liu, Z., Gu, Z., Wang, B., Ouyang, L., Zhao, Z., Chu, T., He, T., Wu, F., Zhang, Q., et al.: Mineru2. 5: A decoupled vision-language model for efficient high-resolution document parsing. arXiv preprint arXiv:2509.22186 (2025)
26. Ouyang, L., Qu, Y., Zhou, H., Zhu, J., Zhang, R., Lin, Q., Wang, B., Zhao, Z., Jiang, M., Zhao, X., et al.: Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24838–24848 (2025)
27. Pasupat, P., Liang, P.: Compositional semantic parsing on semi-structured tables. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1470–1480 (2015)
28. Peng, S., Chakravarthy, A., Lee, S., Wang, X., Balasubramaniyan, R., Chau, D.H.: Unitable: Towards a unified framework for table recognition via self-supervised pretraining. arXiv preprint arXiv:2403.04822 (2024)
29. Prasad, D., Gadpal, A., Kapadni, K., Visave, M., Sultanpure, K.: Cascadetabnet: An approach for end to end table detection and structure recognition from image-based documents. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 572–573 (2020)
30. RapidAI: Rapidtable (2024), `https://github.com/RapidAI/RapidTable`, accessed: 2025-9-25
31. rednote: dots.ocr: Multilingual document layout parsing in a single vision-language model (2025), `https://github.com/rednote-hilab/dots.ocr`
32. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
33. Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q., et al.: Vlm-r1: A stable and generalizable r1-style large vision-language model. arXiv preprint arXiv:2504.07615 (2025)
34. Smock, B., Pesala, R., Abraham, R.: Pubtables-1m: Towards comprehensive table extraction from unstructured documents. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4634–4642 (2022)

35. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
36. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025)
37. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., et al.: Kimi-vl technical report. arXiv preprint arXiv:2504.07491 (2025)
38. Verhoeven, F., Magne, T., Sorkine-Hornung, O.: UVDoc: Neural grid-based document unwarping. In: SIGGRAPH ASIA, Technical Papers (2023)
39. Wan, J., Song, S., Yu, W., Liu, Y., Cheng, W., Huang, F., Bai, X., Yao, C., Yang, Z.: Omniparser: A unified framework for text spotting key information extraction and table recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15641–15653 (2024)
40. Wang, B., Wu, B., Li, W., Fang, M., Liang, Y., Huang, Z., Wang, H., Huang, J., Chen, L., Chu, W., et al.: Infinity parser: Layout aware reinforcement learning for scanned document parsing. arXiv preprint arXiv:2506.03197 (2025)
41. Wang, B., Xu, C., Zhao, X., Ouyang, L., Wu, F., Zhao, Z., Xu, R., Liu, K., Qu, Y., Shang, F., et al.: Mineru: An open-source solution for precise document content extraction. arXiv preprint arXiv:2409.18839 (2024)
42. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
43. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
44. Wei, H., Liu, C., Chen, J., Wang, J., Kong, L., Xu, Y., Ge, Z., Zhao, L., Sun, J., Peng, Y., et al.: General ocr theory: Towards ocr-2.0 via a unified end-to-end model. arXiv preprint arXiv:2409.01704 (2024)
45. Wei, H., Sun, Y., Li, Y.: Deepseek-ocr: Contexts optical compression. arXiv preprint arXiv:2510.18234 (2025)
46. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024)
47. Xie, G.W., Yin, F., Zhang, X.Y., Liu, C.L.: Dewarping document image by displacement flow estimation with fully convolutional network. In: International Workshop on Document Analysis Systems. pp. 131–144. Springer (2020)
48. Xie, G.W., Yin, F., Zhang, X.Y., Liu, C.L.: Document dewarping with control points. In: International conference on document analysis and recognition. pp. 466–480. Springer (2021)
49. Yang, F., Hu, L., Liu, X., Huang, S., Gu, Z.: A large-scale dataset for end-to-end table recognition in the wild. Scientific Data **10**(1), 110 (2023)
50. Yang, Z., Tang, J., Li, Z., Wang, P., Wan, J., Zhong, H., Liu, X., Yang, M., Wang, P., Bai, S., et al.: Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21744–21754 (2025)
51. Ye, J., Qi, X., He, Y., Chen, Y., Gu, D., Gao, P., Xiao, R.: Pingan-vcgroup's solution for icdar 2021 competition on scientific literature parsing task b: table recognition to html. arXiv preprint arXiv:2105.01848 (2021)

52. Yu, T., Wang, Z., Wang, C., Huang, F., Ma, W., He, Z., Cai, T., Chen, W., Huang, Y., Zhao, Y., et al.: Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. arXiv preprint arXiv:2509.18154 (2025)
53. Yu, W., Yang, Z., Wan, J., Song, S., Tang, J., Cheng, W., Liu, Y., Bai, X.: Omniparser v2: Structured-points-of-thought for unified visual text parsing and its generality to multimodal large language models. arXiv preprint arXiv:2502.16161 (2025)
54. Zhao, W., Feng, H., Liu, Q., Tang, J., Wu, B., Liao, L., Wei, S., Ye, Y., Liu, H., Zhou, W., et al.: Tabpedia: Towards comprehensive visual table understanding with concept synergy. Advances in Neural Information Processing Systems **37**, 7185–7212 (2024)
55. Zhao, Y., Huang, J., Hu, J., Wang, X., Mao, Y., Zhang, D., Jiang, Z., Wu, Z., Ai, B., Wang, A., et al.: Swift: a scalable lightweight infrastructure for fine-tuning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 29733–29735 (2025)
56. Zheng, M., Feng, X., Si, Q., She, Q., Lin, Z., Jiang, W., Wang, W.: Multimodal table understanding. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 9102–9124 (2024)
57. Zheng, X., Burdick, D., Popa, L., Zhong, X., Wang, N.X.R.: Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 697–706 (2021)
58. Zhong, X., ShafieiBavani, E., Jimeno Yepes, A.: Image-based table recognition: data, model, and evaluation. In: European conference on computer vision. pp. 564–580. Springer (2020)
59. Zhou, B., Gao, Z., Wang, Z., Zhang, B., Wang, Y., Chen, Z., Xie, H.: Syntab-llava: Enhancing multimodal table understanding with decoupled synthesis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24796–24806 (2025)
60. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)