Multimodal Table Understanding with Difficulty-aware Reinforcement Learning

Chaohu Liu, Haoyu Cao, YongXiang Hua, Linli Xu*

University of Science and Technology of China
State Key Laboratory of Cognitive Intelligence
{liuchaohu, caohaoyu, yx15333063290}@mail.ustc.edu.cn, linlixu@ustc.edu.cn

Abstract

Multimodal table understanding, which aims for a comprehensive grasp of table content by integrating cellular text, tabular structure, and visual presentation, remains a core yet challenging area of research. We identify that the structural complexity of a table, quantifiable by intrinsic properties such as the ratio of merged cells and the total number of cells, presents a significant obstacle for existing models. Our empirical analysis reveals that the performance of leading Multimodal Large Language Models (MLLMs) deteriorates markedly as table complexity increases, exposing a critical vulnerability in their ability to perceive and reason over intricate tabular data. To address this challenge, we propose **MM-Table-R1**, a model enhanced through difficulty-aware reinforcement learning (RL) post-training strategy. Specifically, we introduce both task-level and data-level curriculum learning. The task-level curriculum is designed to establish a capability ladder, where the model first learns basic perceptual and semantic alignment of table data, and then progresses to acquiring multi-step reasoning capabilities. The data-level curriculum ensures that the model is not exposed to difficult samples prematurely, facilitating a more gradual and effective learning process. Furthermore, we invest considerable effort in constructing a high-quality, large-scale training corpus by curating and processing data from diverse open-source table datasets, ensuring that each instance is paired with an objectively verifiable reward signal. Demonstrating exceptional parameter efficiency, our 3B-parameter model sets a new benchmark by surpassing both established 3B and 7B models, including those specifically designed for table reasoning.

1 Introduction

Multimodal table understanding (MTU) involves comprehending and reasoning about data by integrating textual content, structural layout, and visual cues such as fonts, colors, and alignment (Zheng et al. 2024; Lu et al. 2025; Zhang et al. 2025b; Cheng et al. 2025). This process requires models to not only understand semantic information but also interpret complex visual and spatial formatting, enabling them to perform tasks such as Table Question Answering (Pasupat and Liang 2015; Wu et al. 2025), Table Fact Verification (Chen et al. 2019), and Table Reconstruction (Baek et al. 2023).

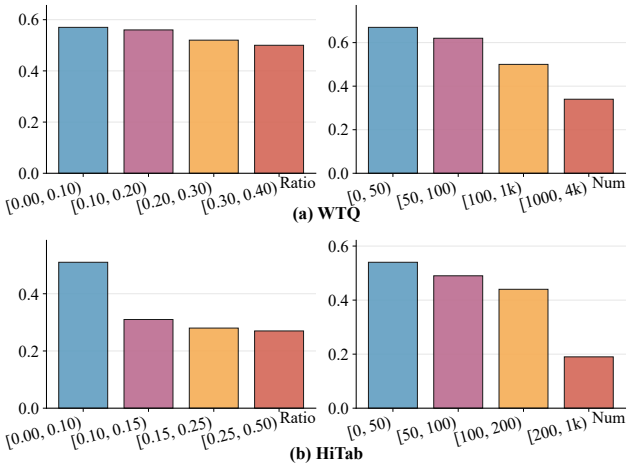

Figure 1: Performance of Qwen2.5-VL in relation to the ratio of merged cells and the total number of cells. Both the WTQ and HiTab datasets contain instances with merged cells. For the WTQ dataset, model accuracy is evaluated based on the original question-answer pairs. For the HiTab dataset, model performance is assessed by the accuracy of table reconstruction.

The widespread presence of tabular data in essential documents, ranging from financial reports and scientific publications to web pages, establishes multimodal table understanding as a critical research area with substantial real-world applications (Tian et al. 2025; Zhang et al. 2025a).

Previous approaches to multimodal table understanding can be broadly categorized into multi-stage and end-to-end methods. Multi-stage methods (Wang et al. 2024; Li et al. 2024; Lei et al. 2025) initially utilize tools such as Optical Character Recognition (OCR) to extract a structured representation (*e.g.*, Markdown, HTML) from a table image, which is then processed by a Large Language Model (LLM) for analysis. The performance of these methods is heavily dependent on the accuracy of the OCR step, making them prone to cumulative errors and often computationally expensive. In contrast, end-to-end approaches that leverage Multimodal Large Language Models (MLLMs) to directly process raw table images represent a more promising di-

*Corresponding Author.

rection due to their enhanced adaptability (Ye et al. 2023; Liu et al. 2024d, 2023). Early works in this domain, such as Table-LLaVA (Zheng et al. 2024) and TabPedia (Zhao et al. 2024), focused on improving visual perception through Supervised Fine-Tuning (SFT). More recent models, such as Turbo (Jiang et al. 2025), have incorporated Reinforcement Learning (RL) to enhance multi-step reasoning capabilities.

However, a critical yet largely overlooked issue is the pronounced sensitivity of contemporary MLLMs to the inherent structural complexity of tables. We propose to quantify this complexity using two key intrinsic metrics: *the ratio of merged cells* and *the total number of cells* (detailed in Section 3). Our empirical analysis substantiates this vulnerability. As illustrated in Figure 1, when evaluating Qwen2.5-VL (Bai et al. 2025) on the WTQ (Pasupat and Liang 2015) question-answering and HiTab (Zhu et al. 2021) table reconstruction benchmarks, we observe a significant performance degradation that directly correlates with an increase in table complexity. We diagnose this as a failure in visuo-structural perception when encountering intricate table layouts. This perceptual deficit creates a critical bottleneck for training, particularly within a reinforcement learning context (Bengio et al. 2009). A standard, uniform sampling strategy lacks a structured curriculum and often confronts the learning agent with highly complex tables prematurely. This approach can overwhelm the agent, leading to training instability and impeding the development of the foundational skills necessary for a robust and generalizable policy (Yang et al. 2025b; Qiu 2025; Yang et al. 2025a).

To address the aforementioned issues, we introduce **MM-Table-R1**, a model enhanced with a difficulty-aware reinforcement learning post-training strategy. Specifically, we incorporate both task-level and data-level curriculum learning. The task-level curriculum aims to build a progressive capability ladder, guiding the model to first *perceive tables* before advancing to *reasoning over tables*. In the first stage, we introduce table reconstruction tasks to help the model develop basic perceptual and semantic alignment skills for table data. In the second stage, we employ verifiable answer-based rewards to stimulate the model's multi-step reasoning ability. The data-level curriculum ensures that the model is not prematurely exposed to excessively difficult samples, thereby fostering a more gradual and stable learning process. This difficulty is determined by both the inherent properties of the table and the model's foundational performance.

To support this framework, we collected, curated, rendered, and filtered a large-scale dataset of high-quality training pairs with verifiable rewards from public benchmarks (Lu et al. 2022; Zheng et al. 2024). Our experiments demonstrate that MM-Table-R1, built upon a 3B parameter backbone, achieves substantial improvements across a variety of table understanding tasks, including WTQ (Pasupat and Liang 2015), TabMWP (Lu et al. 2022), HiTab (Cheng et al. 2021), TAT-QA (Zhu et al. 2021), TabFact (Chen et al. 2019), InfoTabs (Gupta et al. 2020), TableVQA-Bench (Kim, Yim, and Song 2024), and more. Our 3B-parameter MM-Table-R1 showcases exceptional parameter efficiency, establishing a new state-of-the-art by outperforming established 3B and even 7B models, including those specifically architected for table reasoning.

In summary, our main contributions are as follows:

- We empirically demonstrate that intrinsic table properties, specifically the ratio of merged cells and the total number of cells, are significant predictors of performance degradation in Multimodal Large Language Models (MLLMs), highlighting a critical vulnerability in their ability to process complex table structures.
- We introduce **MM-Table-R1**, a model enhanced with a difficulty-aware reinforcement learning (RL) post-training strategy, which integrates both task-level and data-level curriculum learning to effectively address the challenges posed by complex table data.
- We construct a large-scale, high-quality dataset with verifiable rewards, curated from diverse public corpora, which serves as a valuable resource for training advanced table understanding models.
- Through extensive experiments, we demonstrate that our 3B-parameter MM-Table-R1 model achieves state-of-the-art performance, outperforming both established 3B models and even larger 7B models, including those specifically designed for table reasoning tasks.

2 Related Work

Table Understanding

Table Understanding (TU) has become a key task in natural language processing and information extraction, evolving significantly in recent years (Shigarov 2023).

Early table understanding methods, including many modern approaches based on LLMs, were built on a core assumption: the input table must first be converted into a deterministic textual sequence, such as Markdown or HTML format. In this paradigm, models perform tasks such as question answering and fact-checking by processing linearized text (Wu et al. 2025; Jin et al. 2025; Lei et al. 2025; Li et al. 2024; Su et al. 2024; Borisova et al. 2025). However, in many practical scenarios, obtaining high-quality textual representations of tables is highly challenging. Tables in real-world documents are often embedded as images in PDF reports, scanned documents, or webpage screenshots, making their accurate extraction a lossy task. More importantly, these methods struggle to capture complex visual layouts, hierarchical structures, and non-textual elements, which are crucial for comprehensive table understanding.

To overcome the limitations of text-based approaches, Multimodal Table Understanding (MTU) has emerged (Zheng et al. 2024; Singh, Biemann, and Strich 2025). In MTU, models are tasked with generating accurate responses to a variety of table-related queries based on table images. While the accessibility of data in real-world scenarios makes MTU increasingly promising, it also introduces two significant challenges: (1) **Correct Perception of Multimodal Information**: Models must accurately interpret the table's structure, recognize relationships between textual and visual elements, and effectively integrate cross-modal information; and (2) **Complex Multi-step Reasoning**: MTU extends beyond simple information

retrieval, requiring models to exhibit multi-step logical and numerical reasoning capabilities. These challenges remain at the core of ongoing research in the field, and they serve as the motivation behind our design of the task-level curriculum learning approach (Yang et al. 2025a).

Multimodal Large Language Models

Multimodal Large Language Models represent a significant evolution in artificial intelligence, extending the capabilities of traditional Large Language Models (LLMs) beyond text to process and reason over a diverse range of data modalities, including images, audio, and video (Zhang, Chen, and Zhang 2025; Hua et al. 2025). This integration enables a more comprehensive and human-like understanding of complex, real-world information, thereby facilitating their application across a wide array of vision-language tasks, such as detailed image captioning (Liu et al. 2024c; Chen et al. 2024), visual question answering (Bai et al. 2025), and visual document understanding (Liu et al. 2024b; Cao et al. 2023).

A common approach in MTU is adapting general-purpose MLLMs to the specific needs of table-related tasks. For example, Table-LLaVA uses a table-centric dataset for supervised fine-tuning, significantly enhancing the LLaVA model's table understanding capabilities (Zheng et al. 2024). TabPedia introduces a *concept synergy* mechanism that improves the alignment of semantic and structural cues from source-perceived embeddings, boosting the model's comprehension of table content (Zhao et al. 2024). HIPPO employs a modality-consistent strategy for sampling model responses from hybrid-modal tables, enhancing diversity and reducing modality bias during preference optimization training (Liu et al. 2025b; Rafailov et al. 2023).

Reinforcement Learning

Reinforcement Learning (RL) has become one of the most mainstream and powerful techniques in the post-training phase of large models. This method has evolved from early RLHF (Ouyang et al. 2022) for aligning human preferences to the later GRPO (Shao et al. 2024) algorithm, which significantly improves model reasoning using verifiable rewards, undergoing extensive research. In the multimodal domain, NoisyRollout (Liu et al. 2025a) adds noise to images during training to enhance robust image understanding, R1-ShareVL (Yao et al. 2025) performs semantic alignment transformations on prompts to improve their adaptability, and Point-RFT (Ni et al. 2025) leverages external models to describe specific regions of images.

Although reinforcement learning can enhance a model's reasoning abilities, it often faces issues with training instability. Curriculum learning (Bengio et al. 2009) is a common approach to address this problem. It guides the agent's training by designing a sequence of tasks or environments that gradually increase in difficulty (Zhang et al. 2025c).

The work most closely related to ours is Turbo (Jiang et al. 2025), which distills CoT data from DeepSeek-R1 (Guo et al. 2025) for table question answering and uses SFT to bridge the input modality gap. In contrast, our approach integrates RL throughout the entire post-training phase, from perception to reasoning. Moreover, we introduce a novel data-level curriculum learning framework to guide the entire RL training process, ensuring stable learning and capability enhancement on complex tables.

3 Method

This section provides an overview of the approach developed for MM-Table-R1, with the structure shown in Figure 2. We begin by discussing our motivation, followed by the background of reinforcement learning. Finally, we present our task-level and data-level curriculum learning strategies, as well as the construction of our dataset.

Motivation

Multimodal table understanding directly processes visual inputs, enabling the interpretation of rich structural and semantic information that is often lost in text serialization. However, we empirically observed that as tables became more complex, the model's performance significantly declined, suggesting that the model is highly sensitive to the inherent complexity of the tables.

To quantify table complexity, we introduce two metrics derived from its structural properties: the ratio of merged cells and the total number of cells. Let a table T be represented by a set of $\mathcal{L}$ cells. For each cell i, let c_i, m_i and n_i denote its content, rowspan and colspan, respectively. For a non-merged cell, $m_i = n_i = 1$. First, we define the **theoretical maximum number** of cells in the table as $\mathcal{N} = \sum_i m_i \cdot n_i$, which represents the total number of fundamental grid cells the table would occupy if unrolled. Based on this, we define the ratio of merged cells, $\mathcal{R}$, as:

$$\mathcal{R} = \frac{\sum_i^{\mathcal{L}} (m_i \cdot n_i - 1)}{\sum_i^{\mathcal{L}} m_i \cdot n_i} = 1 - \frac{\mathcal{L}}{\mathcal{N}} \tag{1}$$

where a value of $\mathcal{R} = 0$ indicates a table with no merged cells, while a value approaching 1 signifies a highly complex structure dominated by merged cells. Geometrically, this metric quantifies the proportion of the table's area that is occupied by merged cells. For our second metric, we directly adopt the theoretical maximum number of cells $\mathcal{N}$, because it more faithfully captures the true visuo-structural complexity and perceptual load that the table presents.

In our experiments, we selected two test sets with merged cells, WTQ (Pasupat and Liang 2015) and HiTab (Cheng et al. 2021). As shown in Figure 1, the performance of Qwen2.5-VL declines as table complexity increases. This finding highlights a critical bottleneck: models struggle with the foundational perception of complex table structures, and confronting them with such difficult examples prematurely may render the training process inefficient and unstable. This insight is the primary motivation for our methodology. We argue that instead of a standard, mixed-difficulty training regimen, a structured curriculum is necessary. This approach becomes not merely beneficial but essential when leveraging reinforcement learning for post-training. An RL agent exposed to overwhelmingly difficult tasks from the outset will receive sparse and uninformative rewards, leading to policy collapse. Therefore, our proposed task-level and data-level

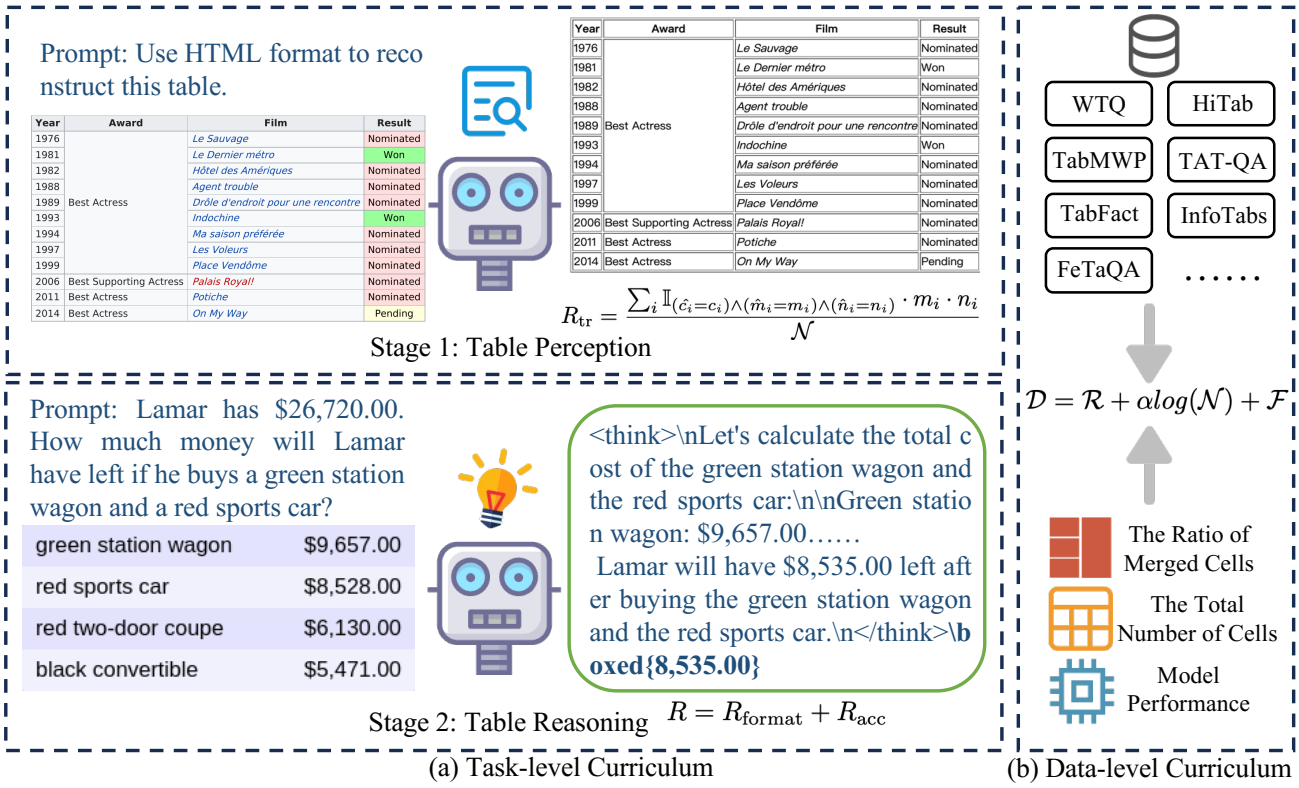

Figure 2: Overview of the difficulty-aware reinforcement learning. It consists of task-level and data-level curriculum learning.

curriculum learning provides a requisite structured pathway, enabling the model to first master fundamental table perception before progressively tackling complex reasoning, ensuring a stable and effective learning trajectory.

Group Relative Policy Optimization

We first provide a formal overview of the reinforcement learning framework. Inspired by recent advances in reasoning models, our approach utilizes the Group Relative Policy Optimization (GRPO) algorithm (Shao et al. 2024). The key innovation of GRPO is its elimination of the conventional critic model, a component often comparable in size to the policy model itself. Instead, GRPO directly estimates the reward baseline from group-level scores, removing the need for a separate critic network. Formally, let q denote a query and $\{o_i\}_{i=1}^{G}$ be a set of outputs sampled from the old policy π_{old}, GRPO maximizes the following objective:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{q\sim P(\mathbb{Q}),\{o_i\}\sim\pi_{\theta_{old}}} \left[\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(r_i(\theta)A_i, \text{clip}\left(r_i(\theta), 1-\varepsilon, 1+\varepsilon\right)A_i\right) - \beta \mathrm{D}_{KL}\left(\pi_\theta||\pi_{ref}\right)\right)\right] \tag{2}$$

$$r_i(\theta) = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)} \tag{3}$$

where ϵ is a hyper-parameter, β is the coefficient controlling the KL divergence penalty term, and π_θ and π_{ref} denote the current and reference policies, respectively. A_i represents the normalized score indicating the relative quality of the i-th response:

$$A_i = \frac{R_i - \text{mean}(\{R_1, \ldots, R_G\})}{\text{std}(\{R_1, \ldots, R_G\})} \tag{4}$$

where R_i is the reward of i-th response.

Task-level Curriculum Learning

Our task-level curriculum learning follows a paradigm that first guides the model to *perceive tables*, and then to perform *reasoning over tables*.

Stage 1 involves the table reconstruction task, which aims to enhance the model's fundamental ability to perceive and understand table data. Unlike previous approaches that utilize SFT (Jiang et al. 2025), this stage incorporates reinforcement learning. This decision arises from the recognition that table reconstruction is not merely a simple OCR task; it requires the model to develop a comprehensive understanding of spatial layouts, attribute relationships, and, critically, the logic of merged cells. This task demands advanced visual reasoning, making RL a more effective paradigm than SFT to develop these reasoning capabilities. On the other hand, SFT generally leads to a degradation in the model's out-of-domain ability, which is detrimental to subsequent reasoning tasks (Chu et al. 2025).

While existing table datasets often employ formats such as Markdown, HTML, or CSV, we choose the HTML format. This decision is based on three key advantages: (1) HTML effectively captures the structure of merged cells in tables through the use of colspan and rowspan attributes, (2) it is widely used in web data, and (3) leading multimodal models, such as Qwen2.5-VL, use the HTML format during their pretraining phase, enabling our post-training process to leverage its inherent table understanding capabilities.

To incentivize the model to capture fine-grained table content, we introduce **Table Verification Reward**. Specifically, we extract the information of each cell in an ordered manner from the sampled response, including the predicted cell content $\hat{c}_i$, rowspan $\hat{m}_i$, and colspan $\hat{n}_i$. These predictions are then compared against the ground truth on a cell-by-cell basis to compute the proportion of correct predictions, which serves as the table reconstruction reward R_{tr}:

$$R_{\text{tr}} = \frac{\sum_i \mathbb{I}_{(\hat{c}_i=c_i)\wedge(\hat{m}_i=m_i)\wedge(\hat{n}_i=n_i)} \cdot m_i \cdot n_i}{\mathcal{N}} \tag{5}$$

where $\mathbb{I}$ is an indicator function that equals 1 only when all conditions are satisfied. This reward formulation provides fine-grained feedback by assigning higher rewards to correctly predicted merged cells, weighted according to their occupied area. The theoretical maximum number of cells $\mathcal{N}$ ensures that the reward is normalized between 0 and 1.

Stage 2 involves regular GRPO training, which primarily enhances the model's multi-step reasoning capabilities through explicit prompting strategies. Specifically, we employ an instruction template that guides the model to perform internal reasoning before generating its final answer: "You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within `<think> </think>` tags. The final answer MUST BE put in `\boxed{}`". This structured prompting encourages the model to explicitly separate reasoning from answer generation, leading to more interpretable and verifiable outputs during reinforcement learning. Accordingly, the overall reward consists of two key components: the format reward R_{format} and the accuracy reward R_{acc}, which measures the correctness of the model's final prediction:

$$R = R_{\text{format}} + R_{\text{acc}} \tag{6}$$

This two-stage training framework equips the model with the ability to interpret table structures effectively and to perform accurate reasoning for complex table-centric tasks.

Data-level Curriculum Learning

GRPO inherently struggles when the model encounters overly difficult samples: a full batch of incorrect responses leads to zero estimated advantage, blocking effective gradient updates. In our preliminary experiments under random sampling (Figure 3), we observed that the model initially failed to make progress, which we attribute to early exposure to high-complexity tables that disrupted its ability to gradually adapt to the task.

To this end, we apply a data-level curriculum learning strategy that organizes training data by a progressively increasing difficulty metric, promoting more stable model convergence. The difficulty of each sample is defined by two factors: (1) the intrinsic properties of the sample, such as the ratio of merged cells and the total number of cells, and (2) the difficulty of samples relative to the current model's performance. For each sample, the model generates 8 answers, and we compute the failure rate $\mathcal{F}$ accordingly. Furthermore, we filter out samples where all answers are correct, as excessively simple samples can also result in a zero-advantage scenario. Formally, the overall difficulty of a sample is defined as follows:

$$\mathcal{D} = \mathcal{R} + \alpha log(\mathcal{N}) + \mathcal{F} \tag{7}$$

where α is a hyperparameter. The logarithm of $\mathcal{N}$ is applied because the complexity growth driven by quantity is non-linear. For example, the increase in perceptual difficulty when the cell number grows from 10 to 100 is different from when it grows from 1000 to 1100. This simple approach enables direct numerical alignment.

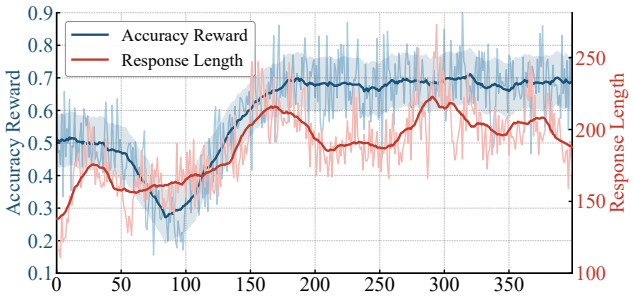

Figure 3: Training curve under random sampling. The model's reward initially decreases before increasing again, indicating instability in the training process.

Data Construction

We construct our training dataset entirely from open-source data, ensuring that each sample contains a table image, the corresponding HTML sequence, and a question-answer pair. The data sources include TabMWP (Lu et al. 2022), WTQ (Pasupat and Liang 2015), HiTab (Cheng et al. 2021), TAT-QA (Zhu et al. 2021), FeTaQA (Nan et al. 2022), TabFact (Chen et al. 2019), InfoTabs (Gupta et al. 2020), and ToTTo (Parikh et al. 2020). For datasets originally provided in CSV or Markdown formats, we customize and apply rule-based scripts to convert tables into HTML format, ensuring content consistency. For datasets without table images, we render them following procedures from prior work (Zheng et al. 2024; Ye et al. 2023). In cases like FeTaQA, where questions and final answers cannot be directly aligned, we use DeepSeekV3 (Liu et al. 2024a) to separate the corresponding question-answer pairs.

4 Experiments

Experimental Setup

Implementation Details. We construct MM-Table-R1 on Qwen2.5-VL-3B. For each stage, we perform reinforcement learning training with a batch size of 256 for two epochs.

Method	Question Answering					Fact Verification		Res.
	TabMWP	WTQ	HiTab	TAT-QA	FeTaQA	TabFact	InfoTabs	ToTTo
InternVL-2.5-8B	90.88	43.19	45.94	34.97	48.30	66.46	55.50	36.34
MiniCPM-V-2.6-8B	83.68	47.97	56.53	51.55	58.30	78.48	73.03	54.88
HIPPO-8B	87.34	55.71	63.13	61.40	59.43	82.29	75.70	-
HIPPO-8B w/o ST	85.83	49.10	57.23	62.22	57.73	80.20	72.74	58.00
Table-LLaVA-7B	53.20	16.62	7.87	10.49	20.78	57.62	66.78	8.76
TabPedia-7B	10.66	23.53	6.54	13.08	19.08	35.49	2.43	2.15
SynTab-LLaVA-7B	88.30	39.59	35.66	51.94	-	70.78	69.42	-
Ovis2-8B	92.00	58.76	68.59	47.67	61.64	80.80	74.11	60.01
Ovis2-CoT-8B	92.12	60.80	66.43	48.70	62.98	81.61	72.46	63.21
Qwen2.5-VL-3B	82.15	56.60	63.42	55.50	57.25	72.15	56.40	48.61
Qwen2.5-VL-7B	92.48	65.85	67.09	70.54	59.06	83.01	77.91	51.92
Turbo-8B	96.75	67.80	72.15	73.21	-	85.81	81.89	-
MM-Table-R1-3B	**97.10**	**74.58**	**76.14**	**74.52**	**69.53**	**87.13**	**85.31**	**78.67**

Table 1: Comprehensive performance comparison of MM-Table-R1-3B against contemporary state-of-the-art models across 8 table understanding benchmarks. The tasks span question answering, fact verification, and table reconstruction (Res.). Our 3B parameter model sets new leading records on all benchmarks, consistently surpassing larger models such as Turbo-8B and Qwen2.5-VL-7B. The lack of open-source access to Turbo and SynTab-LLaVA has prevented a fully aligned evaluation.

Method	Fin.	VWTQ	Syn.	VTab.	AVG.
InternVL2.5	95.2	55.0	60.4	73.2	65.6
HIPPO-8B	89.6	48.9	56.8	81.2	62.3
Ovis2-8B	92.4	59.6	62.4	84.8	69.7
Base-3B	96.0	51.4	59.6	73.6	63.9
Base-7B	97.6	58.5	66.8	81.6	70.2
ours	98.0	69.2	75.6	90.4	78.6

Table 2: The zero-shot evaluation results on the held-out TableVQA-Bench. Our model achieves a significant lead, demonstrating its strong generalization ability.

We use the AdamW optimizer with a learning rate of 1e-6 and a weight decay of 1e-2. Each prompt is sampled with 8 candidate responses, and the maximum response length is 2000. The KL penalty term β is set to 0.01, and α is set to 0.5 by default.

Evaluation Benchmarks. Following prior work (Liu et al. 2025b; Jiang et al. 2025), we evaluate our model's table reasoning ability across a suite of public benchmarks, which include both synthetic and real-world data. We primarily use accuracy as the performance metric. Specifically, for the table question answering task, we utilize the TabMWP (Lu et al. 2022), WTQ (Pasupat and Liang 2015), HiTab (Cheng et al. 2021), TAT-QA (Zhu et al. 2021), and FeTaQA (Nan et al. 2022) datasets. In the case of FeTaQA, we employ DeepSeek-V3 (Liu et al. 2024a) to separate the questions from their corresponding final answers. For the table fact verification task, we conduct evaluations using the TabFact (Chen et al. 2019) and InfoTabs (Gupta et al. 2020) datasets. For the table reconstruction (Res.) task, we report the performance on ToTTo (Parikh et al. 2020) using the R_{tr} metric, which represents cell-level accuracy. We also assess held-out performance in a zero-shot setting on TableVQA-Bench (Kim, Yim, and Song 2024).

Comparison Methods. To provide a comprehensive comparison, we primarily evaluate our model against state-of-the-art open-source general-purpose MLLMs, such as Qwen2.5-VL series (Bai et al. 2025), InternVL-2.5 (Chen et al. 2024), MiniCPM-V-2.6 (Yao et al. 2024), and Ovis2 (Lu et al. 2024), as well as task-specific MLLMs for table-related tasks, including Table-LLaVA (Zheng et al. 2024), TabPedia (Zhao et al. 2024), HIPPO (Liu et al. 2025b), and SynTab-LLaVA (Zhou et al. 2025). We also compare with Turbo (Jiang et al. 2025), a concurrent work, despite its data and models not being released, as it remains a strong reference for multimodal table understanding.

Main Results

To comprehensively evaluate our model's performance, we conduct a comparative analysis against current leading MLLMs on eight widely recognized table understanding benchmarks, with detailed results presented in Table 1. The results clearly demonstrate that our MM-Table-R1-3B model achieves new state-of-the-art performance across all evaluated benchmarks. It is particularly noteworthy that despite being a 3B parameter model, its performance surpasses that of competitors with significantly larger parameter counts, such as Qwen2.5-VL-7B. Especially, Turbo-8B, a model specifically designed for table reasoning, serves as a strong baseline for our method. In Question Answering tasks, MM-Table-R1-3B demonstrates a comprehensive lead. On benchmarks like TabMWP, it surpasses the previous best model, Turbo-8B, with a score of 97.10. On more challenging datasets such as WTQ and HiTab, our model's advantage is even more pronounced. Furthermore, in Fact Verification tasks, which require precise reasoning, our model also sets new performance records on TabFact and InfoTabs with accuracies of 87.13 and 85.31. Notably, in the

Method	HiTab	TabFact	ToTTo	TableVQA
Base	63.42	72.15	48.61	63.94
+SFT	67.91	79.78	65.33	62.73
+S1	66.83	73.27	78.71	64.33
+S2	71.21	83.53	55.32	72.06
+SFT+S2	72.13	82.19	63.61	73.93
ours	76.14	87.13	78.67	78.60

Table 3: Ablation study on different training strategies. We use S1 and S2 to represent Stage 1 and Stage 2, respectively.

Table Reconstruction task, MM-Table-R1-3B outperforms the previous best model by a significant margin, demonstrating its strong capabilities in understanding table structure and content generation.

As shown in Table 2, we also perform a zero-shot evaluation on the held-out TableVQA-Bench dataset, aiming to assess its generalization ability. Our model achieved optimal performance across the average scores, demonstrating exceptional generalization capability. These zero-shot results on unseen datasets strongly validate that our model not only excels on standard benchmarks but also exhibits strong generalization ability, effectively transferring learned knowledge to new challenges.

Ablation Study

The impact of training strategies on performance. We evaluated the model's performance under four scenarios: using only SFT, performing only Stage 1 training, performing only Stage 2 training, and replacing Stage 1 with SFT. As shown in Table 3, the SFT training strategy yields significant performance gains on in-domain tasks but shows a slight decline in performance on out-of-domain tasks. This suggests that while SFT enhances the model's ability to fit known task patterns, it may sacrifice some generalization ability, leading to reduced adaptability in new domains. Next, Stage 1 training alone also results in a modest improvement in the model's overall reasoning performance. This phenomenon demonstrates that enhancing the model's table perception ability is a crucial prerequisite for improving performance on more complex reasoning tasks. However, using SFT for table perception weakens the model's overall generalization ability, and the improvement in perception capability is also limited. Stage 2 training alone also significantly boosts the model's reasoning ability, further validating the tremendous potential of applying reinforcement learning strategies to table tasks. These findings show that Stage 1 builds a solid table perception foundation, while Stage 2 refines reasoning through RL, with both stages together leading to excellent performance and strong generalization across tasks.

The impact of data-level curriculum learning on performance. To further validate the effectiveness of our proposed data-level curriculum learning strategy, we conducted a series of ablation studies. The results are shown in Figure 4. First, we assessed two simplified strategies: No Curriculum Learning (*i.e.*, training in random order) and No Simple Sample Filtering. The results clearly show that both

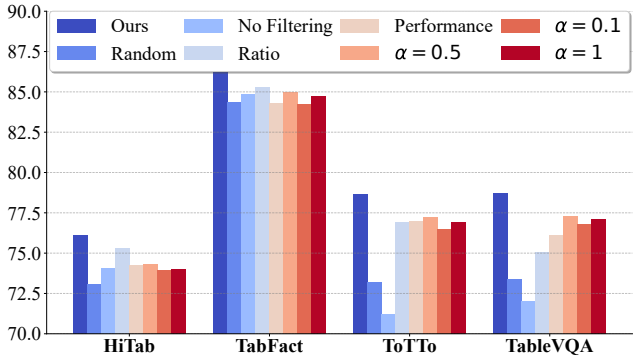

Figure 4: Ablation study on data-level curriculum learning.

approaches led to some performance degradation compared to our full curriculum learning strategy. Next, we explored the independent impact of different difficulty metrics by testing curriculum learning strategies based on a single metric: merged cell ratio, model performance, or total cell count. The results indicate that while each individual metric could guide the model's learning to some extent, none were as effective as our complete strategy that integrates multiple metrics. Building on this, we performed a sensitivity analysis on the hyperparameter α, which weights the cell count metric, testing with $\alpha = 0.1$, $\alpha = 0.5$, and $\alpha = 1.0$. As shown in Figure 4, the model's performance showed low sensitivity to α, indicating good robustness. However, the configuration with $\alpha = 0.5$ achieved the best overall performance, and we therefore adopted it as the default value in our final model.

5 Conclusion

In this paper, we address the challenge of performance degradation in multimodal language models when processing structurally complex tables, proposing MM-Table-R1. The core of this model is an innovative difficulty-aware reinforcement learning strategy that cleverly combines task-level and data-level curriculum learning. The task-level curriculum guides the model to follow a "perceive first, then reason" learning path, while the data-level curriculum ensures the stability of the training process by gradually increasing sample difficulty. With this strategy, our 3B-parameter model achieves state-of-the-art performance on multiple benchmarks, even outperforming several advanced 7B-parameter models, demonstrating exceptional parameter efficiency. This work opens a new path for building more powerful and robust multimodal intelligent systems. This work not only proves the effectiveness of curriculum learning for tackling complex visuo-structural tasks but also opens a new path for building more powerful and robust multimodal intelligent systems.

Acknowledgments

This research was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (XDA0490000) and the National Natural Science Foundation of China (62276245).

References

Baek, Y.; Nam, D.; Surh, J.; Shin, S.; and Kim, S. 2023. TRACE: table reconstruction aligned to corner and edges. In *International Conference on Document Analysis and Recognition*, 472–489. Springer.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.

Borisova, E.; Barth, F.; Feldhus, N.; Ahmad, R. A.; Ostendorff, M.; Suarez, P. O.; Rehm, G.; and Möller, S. 2025. Table Understanding and (Multimodal) LLMs: A Cross-Domain Case Study on Scientific vs. Non-Scientific Data. *arXiv preprint arXiv:2507.00152*.

Cao, H.; Bao, C.; Liu, C.; Chen, H.; Yin, K.; Liu, H.; Liu, Y.; Jiang, D.; and Sun, X. 2023. Attention where it matters: Rethinking visual document understanding with selective region concentration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19517–19527.

Chen, W.; Wang, H.; Chen, J.; Zhang, Y.; Wang, H.; Li, S.; Zhou, X.; and Wang, W. Y. 2019. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*.

Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Cheng, M.; Mao, Q.; Liu, Q.; Zhou, Y.; Li, Y.; Wang, J.; Lin, J.; Cao, J.; and Chen, E. 2025. A survey on table mining with large language models: Challenges, advancements and prospects. *Authorea Preprints*.

Cheng, Z.; Dong, H.; Wang, Z.; Jia, R.; Guo, J.; Gao, Y.; Han, S.; Lou, J.-G.; and Zhang, D. 2021. Hitab: A hierarchical table dataset for question answering and natural language generation. *arXiv preprint arXiv:2108.06712*.

Chu, T.; Zhai, Y.; Yang, J.; Tong, S.; Xie, S.; Schuurmans, D.; Le, Q. V.; Levine, S.; and Ma, Y. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Gupta, V.; Mehta, M.; Nokhiz, P.; and Srikumar, V. 2020. INFOTABS: Inference on tables as semi-structured data. *arXiv preprint arXiv:2005.06117*.

Hua, Y.; Cao, H.; Tao, Z.; Li, B.; Wu, Z.; Liu, C.; and Xu, L. 2025. Input Domain Aware MoE: Decoupling Routing Decisions from Task Optimization in Mixture of Experts. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 5110–5119.

Jiang, J.-P.; Xia, Y.; Sun, H.-L.; Lu, S.; Chen, Q.-G.; Luo, W.; Zhang, K.; Zhan, D.-C.; and Ye, H.-J. 2025. Multimodal Tabular Reasoning with Privileged Structured Information. *arXiv preprint arXiv:2506.04088*.

Jin, R.; Xin, Z.; Xie, X.; Li, Z.; Qi, G.; Chen, Y.; Dai, X.; Wu, T.; and Haffari, G. 2025. Table-r1: Self-supervised and Reinforcement Learning for Program-based Table Reasoning in Small Language Models. *arXiv preprint arXiv:2506.06137*.

Kim, Y.; Yim, M.; and Song, K. Y. 2024. Tablevqa-bench: A visual question answering benchmark on multiple table domains. *arXiv preprint arXiv:2404.19205*.

Lei, F.; Meng, J.; Huang, Y.; Chen, T.; Zhang, Y.; He, S.; Zhao, J.; and Liu, K. 2025. Reasoning-table: Exploring reinforcement learning for table reasoning. *arXiv preprint arXiv:2506.01710*.

Li, P.; He, Y.; Yashar, D.; Cui, W.; Ge, S.; Zhang, H.; Rifinski Fainman, D.; Zhang, D.; and Chaudhuri, S. 2024. Table-gpt: Table fine-tuned gpt for diverse table tasks. *Proceedings of the ACM on Management of Data*, 2(3): 1–28.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, C.; Yin, K.; Cao, H.; Jiang, X.; Li, X.; Liu, Y.; Jiang, D.; Sun, X.; and Xu, L. 2024b. Hrvda: High-resolution visual document assistant. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15534–15545.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024c. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26296–26306.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, X.; Ni, J.; Wu, Z.; Du, C.; Dou, L.; Wang, H.; Pang, T.; and Shieh, M. Q. 2025a. Noisyrollout: Reinforcing visual reasoning with data augmentation. *arXiv preprint arXiv:2504.13055*.

Liu, Y.; Yang, B.; Liu, Q.; Li, Z.; Ma, Z.; Zhang, S.; and Bai, X. 2024d. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*.

Liu, Z.; Wang, H.; Li, X.; Xiong, Q.; Yang, X.; Gu, Y.; Yan, Y.; Shi, Q.; Li, F.; Yu, G.; et al. 2025b. Hippo: Enhancing the table understanding capability of large language models through hybrid-modal preference optimization. *arXiv preprint arXiv:2502.17315*.

Lu, P.; Qiu, L.; Chang, K.-W.; Wu, Y. N.; Zhu, S.-C.; Rajpurohit, T.; Clark, P.; and Kalyan, A. 2022. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*.

Lu, S.; Li, Y.; Chen, Q.-G.; Xu, Z.; Luo, W.; Zhang, K.; and Ye, H.-J. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*.

Lu, W.; Zhang, J.; Fan, J.; Fu, Z.; Chen, Y.; and Du, X. 2025. Large language model for table processing: A survey. *Frontiers of Computer Science*, 19(2): 192350.

Nan, L.; Hsieh, C.; Mao, Z.; Lin, X. V.; Verma, N.; Zhang, R.; Kryściński, W.; Schoelkopf, H.; Kong, R.; Tang, X.; et al. 2022. FeTaQA: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10: 35–49.

Ni, M.; Yang, Z.; Li, L.; Lin, C.-C.; Lin, K.; Zuo, W.; and Wang, L. 2025. Point-rft: Improving multimodal reasoning with visually grounded reinforcement finetuning. *arXiv preprint arXiv:2505.19702*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Parikh, A.; Wang, X.; Gehrmann, S.; Faruqui, M.; Dhingra, B.; Yang, D.; and Das, D. 2020. ToTTo: A Controlled Table-To-Text Generation Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1173–1186.

Pasupat, P.; and Liang, P. 2015. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*.

Qiu, Z. 2025. OpenTable-R1: A Reinforcement Learning Augmented Tool Agent for Open-Domain Table Question Answering. *arXiv preprint arXiv:2507.03018*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shigarov, A. 2023. Table understanding: Problem overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(1): e1482.

Singh, A.; Biemann, C.; and Strich, J. 2025. MTabVQA: Evaluating Multi-Tabular Reasoning of Language Models in Visual Space. *arXiv preprint arXiv:2506.11684*.

Su, A.; Wang, A.; Ye, C.; Zhou, C.; Zhang, G.; Chen, G.; Zhu, G.; Wang, H.; Xu, H.; Chen, H.; et al. 2024. Tablegpt2: A large multimodal model with tabular data integration. *arXiv preprint arXiv:2411.02059*.

Tian, J.; Li, L.; Ye, W.; Wang, H.; Wang, L.; Yu, L.; Ren, Z.; Chen, G.; and Zhao, J. 2025. Toward Real-World Table Agents: Capabilities, Workflows, and Design Principles for LLM-based Table Intelligence. *arXiv preprint arXiv:2507.10281*.

Wang, Z.; Zhang, H.; Li, C.-L.; Eisenschlos, J. M.; Perot, V.; Wang, Z.; Miculicich, L.; Fujii, Y.; Shang, J.; Lee, C.-Y.; et al. 2024. Chain-of-table: Evolving tables in the reasoning chain for table understanding. *arXiv preprint arXiv:2401.04398*.

Wu, Z.; Yang, J.; Liu, J.; Wu, X.; Pan, C.; Zhang, J.; Zhao, Y.; Song, S.; Li, Y.; and Li, Z. 2025. Table-r1: Region-based reinforcement learning for table understanding. *arXiv preprint arXiv:2505.12415*.

Yang, B.; Zhang, Y.; Liu, D.; Freitas, A.; and Lin, C. 2025a. Does table source matter? benchmarking and improving multimodal scientific table understanding and reasoning. *arXiv preprint arXiv:2501.13042*.

Yang, Z.; Chen, L.; Cohan, A.; and Zhao, Y. 2025b. Table-r1: Inference-time scaling for table reasoning. *arXiv preprint arXiv:2505.23621*.

Yao, H.; Yin, Q.; Zhang, J.; Yang, M.; Wang, Y.; Wu, W.; Su, F.; Shen, L.; Qiu, M.; Tao, D.; et al. 2025. R1-ShareVL: Incentivizing Reasoning Capability of Multimodal Large Language Models via Share-GRPO. *arXiv preprint arXiv:2505.16673*.

Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Ye, J.; Hu, A.; Xu, H.; Ye, Q.; Yan, M.; Xu, G.; Li, C.; Tian, J.; Qian, Q.; Zhang, J.; et al. 2023. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*.

Zhang, J.; Chen, P.; and Zhang, Y. 2025. TableMoE: Neuro-Symbolic Routing for Structured Expert Reasoning in Multimodal Table Understanding. *arXiv preprint arXiv:2506.21393*.

Zhang, X.; Song, A.; Qiu, J.; Jin, J.; Zhang, T.; and Fang, X. 2025a. Exploring Multimodal Relation Extraction of Hierarchical Tabular Data with Multi-task Learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 26770–26781.

Zhang, X.; Wang, D.; Dou, L.; Zhu, Q.; and Che, W. 2025b. A survey of table reasoning with large language models. *Frontiers of Computer Science*, 19(9): 199348.

Zhang, X.; Wang, J.; Cheng, Z.; Zhuang, W.; Lin, Z.; Zhang, M.; Wang, S.; Cui, Y.; Wang, C.; Peng, J.; et al. 2025c. Srpo: A cross-domain implementation of large-scale reinforcement learning on llm. *arXiv preprint arXiv:2504.14286*.

Zhao, W.; Feng, H.; Liu, Q.; Tang, J.; Wu, B.; Liao, L.; Wei, S.; Ye, Y.; Liu, H.; Zhou, W.; et al. 2024. Tabpedia: Towards comprehensive visual table understanding with concept synergy. *Advances in Neural Information Processing Systems*, 37: 7185–7212.

Zheng, M.; Feng, X.; Si, Q.; She, Q.; Lin, Z.; Jiang, W.; and Wang, W. 2024. Multimodal table understanding. *arXiv preprint arXiv:2406.08100*.

Zhou, B.; Gao, Z.; Wang, Z.; Zhang, B.; Wang, Y.; Chen, Z.; and Xie, H. 2025. SynTab-LLaVA: Enhancing Multimodal Table Understanding with Decoupled Synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24796–24806.

Zhu, F.; Lei, W.; Huang, Y.; Wang, C.; Zhang, S.; Lv, J.; Feng, F.; and Chua, T.-S. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. *arXiv preprint arXiv:2105.07624*.