Can GRPO Boost Complex Multimodal Table Understanding?

Xiaoqiang Kang[1,2], **Shengen Wu**[3], **Zimu Wang**[1,2], **Yilin Liu**[4], **Xiaobo Jin**[1], **Kaizhu Huang**[5], **Wei Wang**[1], **Yutao Yue**[3], **Xiaowei Huang**[2], **Qiufeng Wang**[1,*]
[1]School of Advanced Technology, Xi'an Jiaotong-Liverpool University
[2]Department of Computer Science, University of Liverpool
[3]Information Hub, Hong Kong University of Science and Technology (Guangzhou)
[4]University of Southern California [5]Duke Kunshan University
Xiaoqiang.Kang23@student.xjtlu.edu.cn, Qiufeng.Wang@xjtlu.edu.cn

Abstract

Existing table understanding methods face challenges due to complex table structures and intricate logical reasoning. While supervised fine-tuning (SFT) dominates existing research, reinforcement learning (RL), such as Group Relative Policy Optimization (GRPO), has shown promise but struggled with low initial policy accuracy and coarse rewards in tabular contexts. In this paper, we introduce Table-R1, a three-stage RL framework that enhances multimodal table understanding through: (1) **Warm-up** that prompts initial perception and reasoning capabilities, (2) **Perception Alignment GRPO (PA-GRPO)**, which employs continuous Tree-Edit-Distance Similarity (TEDS) rewards for recognizing table structures and contents, and (3) **Hint-Completion GRPO (HC-GRPO)**, which utilizes fine-grained rewards of residual steps based on the hint-guided question. Extensive experiments demonstrate that Table-R1 can boost the model's table reasoning performance obviously on both held-in and held-out datasets, outperforming SFT and GRPO largely. Notably, Qwen2-VL-7B with Table-R1 surpasses larger specific table understanding models (e.g., Table-LLaVA 13B), even achieving comparable performance to the closed-source model GPT-4o on held-in datasets, demonstrating the efficacy of each stage of Table-R1 in overcoming initialization bottlenecks and reward sparsity, thereby advancing robust multimodal table understanding.

1 Introduction

Table understanding is regarded as a cornerstone task in NLP and multimodal research, as structured data in the form of tables is pervasive across diverse domains such as scientific research (Van Breugel and Van Der Schaar, 2024; Li et al., 2024a), finance (Chen et al., 2021; Katsis et al., 2022), and education (Lu et al., 2023; Kang et al., 2025). This task presents unique challenges due to the complex table structures and intricate logical reasoning in real-world tables (Mathur et al., 2024; Zhao et al., 2024). Effectively interpreting and reasoning over tabular data is critical for enhancing information extraction and automating data analysis.

Recent research on table understanding has witnessed two predominant paradigms: supervised fine-tuning (SFT) and reinforcement learning (RL). While most work has been largely dominated by SFT (Zheng et al., 2024; Kang et al., 2025), these methods suffer from limited generalization when facing unseen table structures or complex reasoning chains (Chu et al., 2025). In contrast, RL has resurged as a promising paradigm for improv-

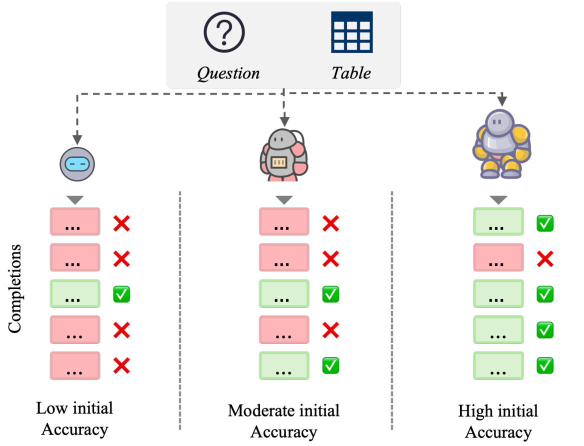


Figure 1: Comparative analysis of different initial policy accuracy in a group.

Model	A_{init} (%)	A_{final} (%)	ΔA	$V_{A_{init}}$
Qwen2.5-0.5B	21.4	28.7	7.3	0.168
Qwen2.5-1.5B	31.2	44.0	12.8	0.215
Qwen2.5-3B	55.2	87.6	**32.4**	**0.247**
Qwen2.5-7B	81.8	91.2	9.4	0.149

Table 1: Comparative analysis of GRPO performance on TabMWP across Qwen2.5 models of varying scales. $\Delta A = A_{\text{final}} - A_{\text{init}}$ represents the absolute improvement in accuracy, while $\text{V}(A_{\text{init}}) = A_{\text{init}}(1 - A_{\text{init}})$ denotes the variance-based measure of initial policy accuracy.

[*] Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 12643–12656
November 4-9, 2025 ©2025 Association for Computational Linguistics

ing complex reasoning, especially in mathematical tasks. Methods such as Proximal Policy Optimization (PPO, Schulman et al., 2017), Direct Preference Optimization (DPO, Rafailov et al., 2023), and Group Relative Policy Optimization (GRPO, Shao et al., 2024) demonstrate that RL-based methods can significantly enhance reasoning capabilities. However, the application of RL to multimodal table understanding remains underexplored, despite its potential to address the limitations of SFT-based approaches. This naturally raises an important research question: *Can RL-based methods such as GRPO be effectively adapted to enhance complex multimodal table understanding for Large Vision Language Models (LVLMs)?*

To accomplish this, we first conduct a preliminary study to investigate the application of GRPO to complex table understanding tasks, identifying a critical dependency on the initial policy's accuracy (see Section 3). As shown in Figure 1 and Table 1, only a policy model with moderate accuracy can produce a balanced mix of correct and incorrect outputs, which is crucial for policy optimization. This finding highlights a fundamental limitation: the low initial accuracy of the policy model hinders effective back propagation due to the low standard of rewards, ultimately impairing the convergence of the policy model. Additionally, existing reward functions primarily depend on binary correctness signals. Thus, another challenge is how to devise more fine-grained reward functions tailored for tabular perception and reasoning tasks.

To address the challenges highlighted, we introduce Table-R1, the first RL-based framework specifically designed for multimodal table understanding. Inspired by the cold-start strategy in DeepSeek-R1 (Guo et al., 2025), Table-R1 introduces a three-stage framework (see Figure 2): (1) **Warm-up** initializes the model with perception and reasoning capabilities, while also boosting the policy model's initial accuracy. (2) **Perception Alignment GRPO (PA-GRPO)** employs continuous reward signals, Tree-Edit-Distance-based Similarity (TEDS), for table structure recognition. (3) **Hint-Completion GRPO (HC-GRPO)** applies reward functions to the residual steps of the hint-guided question, which offers a finer-grained reward than a coarse solution-level reward and further refines the model's reasoning capabilities.

We divide our datasets into two parts: held-in and held-out. The held-in comprises 4 multimodal table understanding tasks for training, whereas similar tasks are set as held-out to assess the model's robustness. We validate the effectiveness of each stage of Table-R1 and conduct comprehensive experiments compared against baselines. Experimental results indicate that Table-R1 consistently outperforms both SFT and GRPO across models of different scales. For Qwen2-VL-7B, Table-R1 achieves a 3.93% improvement over SFT and a 16.38% improvement over GRPO on held-in, as well as a 7.72% improvement over SFT and a 8.79% improvement over GRPO on held-out, significantly surpassing that of models with larger scale (e.g., Table-LLaVA 13B) and matching GPT-4o's performance.

The main contributions of our work are summarized as follows: **(1)** We identify and empirically validate the pivotal limitation of GRPO in table reasoning, that the policy model is sensitive to the initial accuracy; **(2)** We propose Table-R1, a new three-stage reinforcement learning framework that enables LVLM to improve its perception and reasoning capability for the first time; **(3)** We conduct comprehensive experiments on six datasets to demonstrate that our framework can obviously surpass both SFT and GRPO, specifically boosting the Qwen2-VL-7B model largely to achieve state-of-the-art performance on several benchmarks.

2 Related Work

Multimodal Table Understanding is a fundamental task in computer vision and document understanding. Early works have focused on visual table recognition, structure parsing, and content extraction from document images, such as PubTabNet (Zhong et al., 2020), FinTabNet (Zheng et al., 2021), and TableFormer (Yang et al., 2022).

Recent efforts have advanced toward reasoning over visually and contextually rich tables. Representative works include Table-LLaVA (Zheng et al., 2024), which augments table inputs with cell-associated images, and TabPedia (Zhao et al., 2024), which provides a large-scale multimodal table pretraining corpus to improve downstream performance. Multimodal ArXiv (Li et al., 2024a) proposes fine-grained reasoning over scientific tables with linked charts and text, while Karma (Mathur et al., 2024) incorporates symbolic knowledge graphs for better factual alignment. In terms of reasoning supervision, Cheng et al. (2024) proposes R3V, a self-training framework that iteratively generates and selects chain-of-thought trajectories to improve multimodal question answering on documents and tables.

Reinforcement Learning (RL), as a machine learning paradigm, aims to learn optimal decision-making by enabling an agent to interact with an environment and relying on reward signals (Zhang et al., 2025). In the context of large language models (LLMs), RL is mapped to concrete language generation tasks: the LLM functions as the agent, with user prompts and generated text constituting the environment state, while generating the next token corresponds to the agent's action (Wang et al., 2025). To facilitate effective training, pre-trained reward models are typically employed. These models automatically evaluate the quality of the generated text based on human preferences or preset criteria, and their outputs serve as rewards that guide the training of the LLM (Ouyang et al., 2022).

In recent years, RL techniques have substantially enhanced the reasoning capabilities of LLMs (Luong et al., 2024; Liu et al., 2025b; Peng et al., 2025). Numerous studies have adopted appropriate reward functions and policy optimization strategies to reinforce high-quality reasoning paths while penalizing low-quality ones, thereby guiding the models to achieve more coherent and logically structured reasoning trajectories (Wang et al., 2025). For example, Rafailov et al. (2023) and Yuan et al. (2025) employ Direct Preference Optimization (DPO), Zhang et al. (2024) utilizes a process reward model to evaluate each reasoning step, and Wang et al. (2024a) leverages both process and outcome reward models simultaneously. Particularly noteworthy is Deepseek-R1, which employs a Group Relative Policy Optimization (GRPO) method to achieve robust reasoning capabilities solely through RL (Guo et al., 2025). GRPO replaces the traditional reward function with verifiable rule-based rewards and substitutes multiple sampling for the critic model, directly steering the model to converge on high-quality reasoning strategies without the need for complex reward modeling(Shao et al., 2024).

3 Observation

We first investigate the application of GRPO to complex table understanding tasks, identifying a critical dependency on the initial policy's accuracy. As detailed in Table 1, our evaluation on the TabMWP dataset (Lu et al., 2023) reveals a stark performance disparity. Models with either low (e.g., Qwen2.5-0.5B) or high (e.g., Qwen2.5-7B) initial accuracy merely yield gains of 7.3% and 9.4%, respectively. In contrast, the Qwen2.5-3B model, starting from a moderate accuracy, achieves a substantial improvement of 32.4%. This phenomenon is visually represented in Figure 1, where a low-accuracy policy tends to generate mostly incorrect solutions, while a high-accuracy one generates predominantly correct solutions. Only a policy model with moderate accuracy can produce a balanced mix of correct and incorrect outputs, which is crucial for policy optimization.

This behavior can be explained by the variance of the binary reward. Assuming rewards follow a Bernoulli distribution $R \sim \text{Bernoulli}(p)$, where $p = A_{\text{init}}$, the variance is given by $A_{\text{init}}(1 - A_{\text{init}})$. When A_{init} approaches 0 or 1, the variance approaches zero. This leads to a zero advantage estimate and negligible policy gradients for optimization. Conversely, when $A_{\text{init}} \approx 0.5$ (e.g., Qwen2.5-3B at 55.2%), variance approaches the theoretical maximum of 0.25, maximizing advantage and providing strong gradients for effective policy optimization. This finding highlights a fundamental limitation: the low initial accuracy of the policy model hinders effective back propagation due to the low standard of rewards, ultimately impairing the convergence of the policy model. This observation aligns with recent findings emphasizing GRPO's sensitivity to policy initialization (Yu et al., 2025; Liu et al., 2025a).

4 Methodology

4.1 Problem Formulation

The multimodal tabular input, (I, Q), consists of a question Q and a corresponding table image I. The policy model π_θ generates a series of actions to generate token sequences, which comprises a step-by-step reasoning trajectory and the final answer. Each action corresponds to generating the next token in the output sequence. For each rollout, the model produces candidate outputs $\{S^1, \dots, S^G\}$ with corresponding rewards $\{R^1, \dots, R^G\}$. The objective is to optimize π_θ to maximize the expected cumulative reward $\mathbb{E}_{S \sim \pi_\theta(I,Q)}[R(S)]$ by selecting actions that generate high-quality reasoning trajectories.

4.2 Table-R1 Framework

As illustrated in Figure 2, we initially propose Table-R1, a three-stage training framework for tabular perception and reasoning tasks. (1) **Warm-up**: Supervised fine-tuning to initialize the model

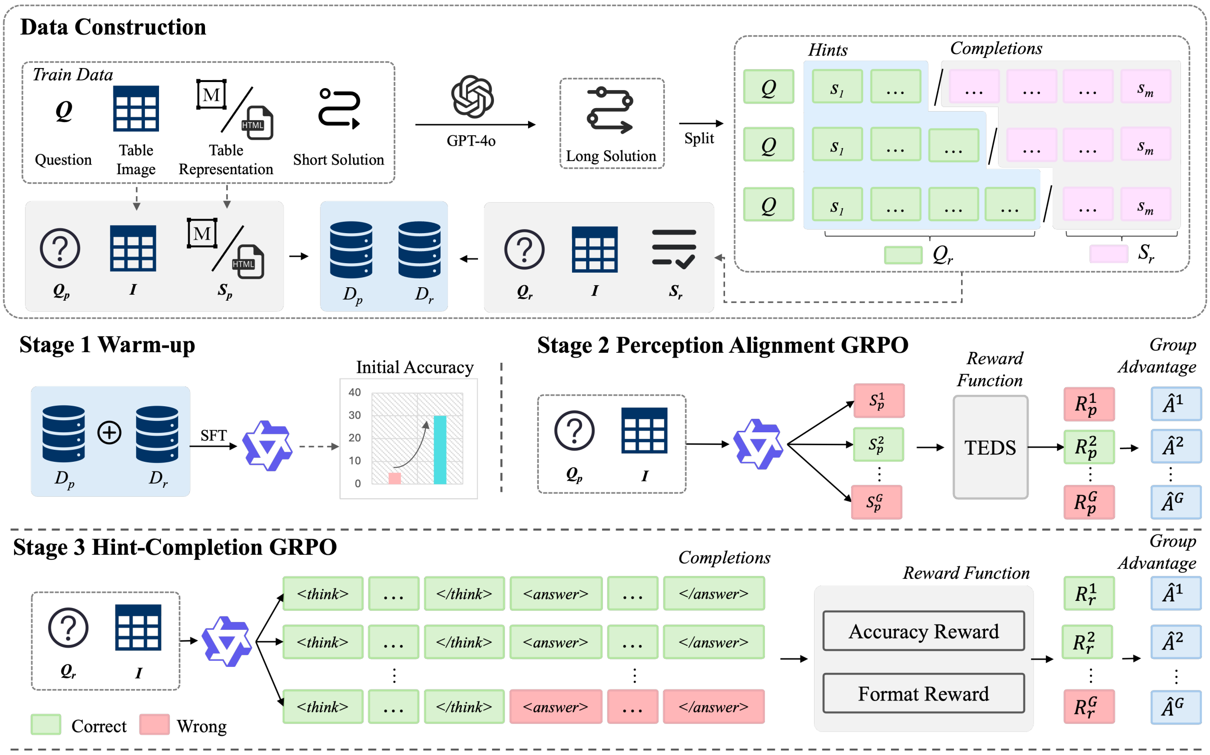


Figure 2: Overall framework of Table-R1. (1) **Warm-up** establishes foundational capabilities in both visual perception and reasoning. (2) **PA-GRPO** refines the model's structural understanding by employing TEDS as a continuous reward. (3) **HC-GRPO** utilizes fine-grained rewards of residual steps based on the hint-guided question.

with strong perception and reasoning capabilities. (2) **Perception Alignment GRPO**: Improves table structure recognition using continuous rewards. (3) **Hint-Completion GRPO**: Enhances step-by-step reasoning through hint-based completions. The overall training algorithm is shown in Appendix A.

Warm-up. As shown in Figure 1, the initial accuracy of the policy model plays a crucial role during the training of GRPO. To address this, we introduce a warm-up stage that significantly boosts the model's initial perception and reasoning accuracy via SFT. This stage equips the policy model with the ability to both convert images to structured table representations and to generate valid step-wise reasoning paths. During the warm-up stage, the policy model undergoes SFT using the perception task dataset D_p and reasoning task dataset D_r, which will be detailed in the following two stages. The loss function we used here is:

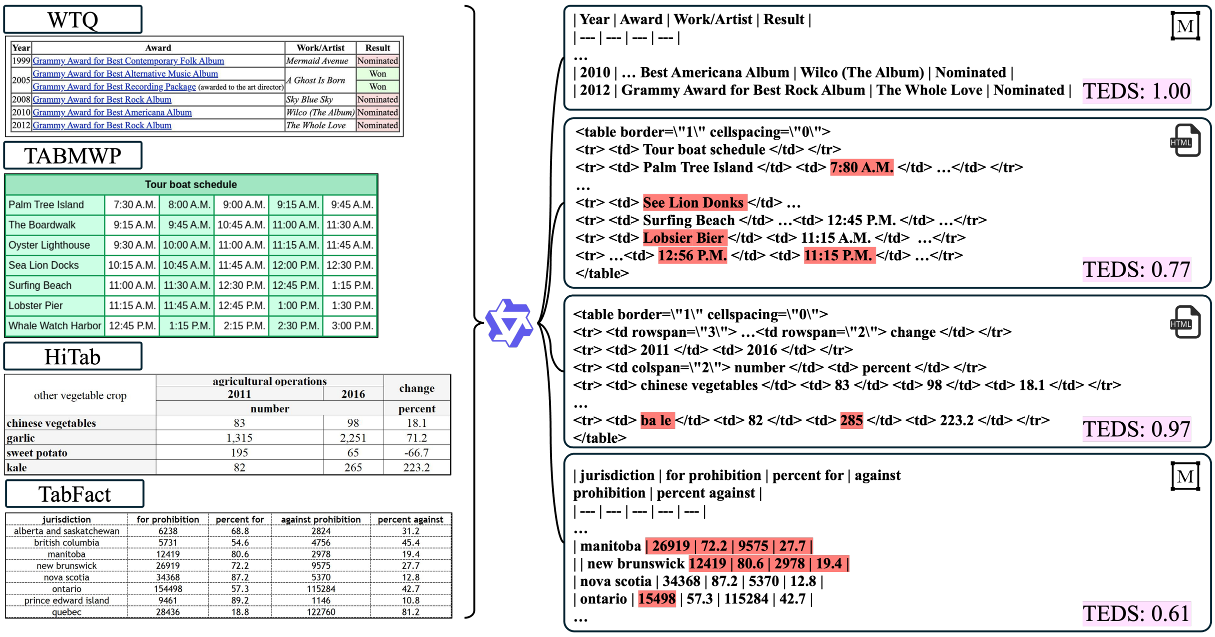
$$\mathcal{L}_{\text{warm-up}} = -\mathbb{E}_{(I,Q,S)\sim D_p \cup D_r}\left[\sum_{t=1}^{T} \log \pi_\theta(s_t|s_{<t})\right]. \tag{1}$$

Perception-Alignment GRPO (PA-GRPO). In this stage, the model focuses on its ability to recognize patterns and structures. The model extracts structured tabular representations from input images I, generating outputs S_p in either Markdown or HTML format. To enhance the linguistic diversity of instruction, we have constructed 20 distinct instruction variants Q_p for this task, as shown in Figure 5 in Appendix B.2. The complete dataset, denoted as D_p, is a collection of tuples (I, Q_p, S_p), where each tuple consists of a table image I, an instruction variant Q_p, and the target structured representation S_p. Tree-Edit-Distance-based Similarity (TEDS) (Zhong et al., 2019) is utilized as a reward. This similarity is calculated based on the tree structure of the table sequence. It assesses both structural similarity and content similarity of the cells between the predicted table S_p and the golden answer GA. TEDS is normalized on a scale from 0 to 1, where a score of 1 indicates a perfect match. Several detailed examples are reported in Figure 3. Formally, the reward is defined as follows:

$$R_p = \text{TEDS}(S_p, GA). \tag{2}$$

Since this perception task doesn't require a reasoning process, LVLM is expected to provide direct answers, and the complete prompt is displayed in Table 10 in Appendix B.1.

Hint-Completion GRPO (HC-GRPO). During this stage, given a question Q, the model enhances



Year	Award	Work/Artist	Result
1999	Grammy Award for Best Contemporary Folk Album	*Mermaid Avenue*	Nominated
2005	Grammy Award for Best Alternative Music Album	*A Ghost Is Born*	Won
	Grammy Award for Best Recording Package (awarded to the art director)		Won
2008	Grammy Award for Best Rock Album	*Sky Blue Sky*	Nominated
2010	Grammy Award for Best Americana Album	*Wilco (The Album)*	Nominated
2012	Grammy Award for Best Rock Album	*The Whole Love*	Nominated

Tour boat schedule					
Palm Tree Island	7:30 A.M.	8:00 A.M.	9:00 A.M.	9:15 A.M.	9:45 A.M.
The Boardwalk	9:15 A.M.	9:45 A.M.	10:45 A.M.	11:00 A.M.	11:30 A.M.
Oyster Lighthouse	9:30 A.M.	10:00 A.M.	11:00 A.M.	11:15 A.M.	11:45 A.M.
Sea Lion Docks	10:15 A.M.	10:45 A.M.	11:45 A.M.	12:00 P.M.	12:30 P.M.
Surfing Beach	11:00 A.M.	11:30 A.M.	12:30 P.M.	12:45 P.M.	1:15 P.M.
Lobster Pier	11:15 A.M.	11:45 A.M.	12:45 P.M.	1:00 P.M.	1:30 P.M.
Whale Watch Harbor	12:45 P.M.	1:15 P.M.	2:15 P.M.	2:30 P.M.	3:00 P.M.

other vegetable crop	agricultural operations		change
	2011	2016	
	number		percent
chinese vegetables	83	98	18.1
garlic	1,315	2,251	71.2
sweet potato	195	65	-66.7
kale	82	265	223.2

jurisdiction	for prohibition	percent for	against prohibition	percent against
alberta and saskatchewan	6238	68.8	2824	31.2
british columbia	5731	54.6	4756	45.4
manitoba	12419	80.6	2978	19.4
new brunswick	26919	72.2	9575	27.7
nova scotia	34368	87.2	5370	12.8
ontario	154498	57.3	115284	42.7
prince edward island	9461	89.2	1146	10.8
quebec	28436	18.8	122760	81.2

```
| Year | Award | Work/Artist | Result |
| --- | --- | --- | --- |
…
| 2010 | … Best Americana Album | Wilco (The Album) | Nominated |
| 2012 | Grammy Award for Best Rock Album | The Whole Love | Nominated |
```

```
<table border=\"1\" cellspacing=\"0\">
<tr> <td> Tour boat schedule </td> </tr>
<tr> <td> Palm Tree Island </td> <td> 7:80 A.M. </td> …</td> </tr>
…
<tr> <td> See Lion Donks </td> …
<tr> <td> Surfing Beach </td> …<td> 12:45 P.M. </td> …</tr>
<tr> <td> Lobsier Bier </td> <td> 11:15 A.M. </td>  …</tr>
<tr> …<td> 12:56 P.M. </td> <td> 11:15 P.M. </td> …</tr>
</table>
```

```
<table border=\"1\" cellspacing=\"0\">
<tr> <td rowspan=\"3\"> …<td rowspan=\"2\"> change </td> </tr>
<tr> <td> 2011 </td> <td> 2016 </td> </tr>
<tr> <td colspan=\"2\"> number </td> <td> percent </td> </tr>
<tr> <td> chinese vegetables </td> <td> 83 </td> <td> 98 </td> <td> 18.1 </td> </tr>
…
<tr> <td> ba le </td> <td> 82 </td> <td> 285 </td> <td> 223.2 </td> </tr>
</table>
```

```
| jurisdiction | for prohibition | percent for | against
prohibition | percent against |
| --- | --- | --- | --- | --- |
…
| manitoba | 26919 | 72.2 | 9575 | 27.7 |
| | new brunswick 12419 | 80.6 | 2978 | 19.4 |
| nova scotia | 34368 | 87.2 | 5370 | 12.8 |
| ontario | 15498 | 57.3 | 115284 | 42.7 |
…
```

Figure 3: Examples from the datasets used in PA-GRPO. The highlighted red segments indicate the incorrect predictions. TEDS assigns a continuous score to each output, reflecting the similarity to the golden answer.

its reasoning capability by progressively completing the remaining steps to reach the final answer. The residual-step rewards can be more fine-grained than solution-level. Some initial solutions are too brief to be effectively split into two parts, so we employ GPT-4o (OpenAI et al., 2024) to expand short solutions into long reasoning chains $[s_1, s_2, \dots, s_n]$, where each s_i represents the i-th step of the extended solution. The detailed prompts are displayed in Figure 6 in Appendix B.3.

For training data generation, each expanded solution is randomly divided into two segments at position $j \sim \text{Uniform}\{1, \dots, m-1\}$. The first segment, called **Hints**, includes the initial reasoning steps and is represented as $[s_1, \dots, s_j]$ The remaining steps $S_r = [s_{j+1}, \dots, s_m]$ are referred to as the **Completions**. These hints, when combined with the original question Q, constitute the input query $Q_r = [Q, s_1, \dots, s_j]$. By default, a long solution can generate three hint-completion pairs. The full dataset for this stage, D_r, is thus composed of tuples (I, Q_r, S_r). For the reasoning task, the LVLM is expected to first perform step-by-step reasoning before generating the final answer. The full prompt is provided in Table 10 in Appendix B.1.

The reward function consists of two components: an accuracy reward R_{acc} and a format reward R_{format}, combined as follows:

$$R_{\text{r}} = R_{\text{acc}} + R_{\text{format}}. \tag{3}$$

Accuracy reward is calculated by comparing the model-generated answer MA, extracted from within `<answer></answer>` tags, with the golden answer GA, using a binary reward scheme:

$$R_{\text{acc}} = \begin{cases} 1, & \text{if } MA = GA \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

To ensure rigorous and consistent assessment, we adopt the widely used Math-Verify[1] library, which provides a standardized method for parsing and verifying mathematical expressions and numerical values when comparing MA and GA. **Format reward** incentivizes the model to organize its output correctly by placing the reasoning within `<think></think>` tags and the final answer within `<answer></answer>` tags. This is checked via regular expression matching (REM):

$$R_{\text{format}} = \begin{cases} 1, & \text{if } REM(S_r) = True \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Unified GRPO-Based Training Objective. For both PA-GRPO and HC-GRPO, we employ a unified policy optimization strategy, differing only in the reward definition. Following Shao et al. (2024), after computing the reward R^i for each output S^i in a rollout, the advantage is calculated as:

$$\hat{A}^i = \frac{R^i - \text{mean}(\{R^1, \dots, R^G\})}{\text{std}(\{R^1, \dots, R^G\})}, \tag{6}$$

[1] https://github.com/huggingface/Math-Verify

Task Category	Task Name	Dataset	Table Style	Source	Held-in	Original		Sampled		Avg. Pixel
						# T	# Q	# T	# Q	
Table Question Answering (TQA)	Flat TQA	WTQ (2015)	W	Wikipedia	Yes	1.6K	17K	1.6K	8K	1992×1116
	Hierarchical TQA	HiTab (2022)	E	Wikipedia Goverment Reports	Yes	3K	8K	3K	8K	3057×793
	Tabular Numerical Reasoning	TabMWP (2023)	W	Math Exams	Yes	30K	30K	8K	8K	267×191
		TAT-QA (2021)	M	Financial Reports	No	1.7K	5.9K	/	/	2446×1141
Table Fact Verification (TFV)	TFV	TabFact (2020)	E, M	Wikipedia	Yes	9K	31K	8K	8K	2440×900
		InfoTabs (2020)	W	Wikipedia	No	1.9K	18K	/	/	792×880

Table 2: Statistics of our constructed train datasets, which are sampled from the original datasets. W, E, and M represent Web page, Excel, and Markdown tables, respectively. The symbols # T and # Q indicate the number of tables and questions, and Avg. means average.

which normalizes reward relative across the group.

The overall loss function combines a clipped surrogate objective with a KL divergence penalty:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathcal{L}_{\text{clip}}(\theta) - \beta\mathbb{D}_{\text{KL}}[\pi_\theta \| \pi_{\text{ref}}], \quad (7)$$

where $\mathcal{L}_{\text{clip}}(\theta)$ adopts the proximal policy optimization mechanism:

$$\mathcal{L}_{\text{clip}}(\theta) = \frac{1}{G}\sum_{i=1}^{G} \min\Bigg(\frac{\pi_\theta(S^i|Q,I)}{\pi_{\theta_{\text{old}}}(S^i|Q,I)}\hat{A}^i, \\ \text{clip}\Big(\frac{\pi_\theta(S^i|Q,I)}{\pi_{\theta_{\text{old}}}(S^i|Q,I)},\ 1{-}\epsilon,\ 1{+}\epsilon\Big)\hat{A}^i\Bigg), \quad (8)$$

constraining policy model updates to prevent destructive parameter changes. The KL divergence term $\mathbb{D}_{\text{KL}}$ regularizes the policy model π_θ to maintain proximity to the reference model π_{ref}, which is initialized as a frozen copy of the pretrained policy model. This dual mechanism balances reward maximization with behavioral consistency, mitigating catastrophic forgetting during GRPO training.

5 Experiments

5.1 Baselines

We compare the performance of Table-R1 against the following baselines: (1) *Open-source LVLMs*: Qwen2-VL (Wang et al., 2024b), Qwen2-VL-7B-Ins (Yang et al., 2025), DeepSeek-VL2 (Wu et al., 2024), LLaVA v1.5 (Liu et al., 2024), Table-LLaVA (Zheng et al., 2024), mPLUG-Owl2 (Ye et al., 2023), Monkey (Li et al., 2024b), and QVQ-72B-Preview (Team, 2024). (2) *Closed-source LVLMs*: GPT-4o (OpenAI et al., 2024), Gemini 2.5 Pro (Team et al., 2023) and Claude-3.5-Sonnet (Anthropic, 2024). (3) Two model optimization methods: SFT and GRPO.

5.2 Datasets and Evaluation

We conduct experiments on MMTab (Zheng et al., 2024), which is a recent large-scale dataset focused on multimodal table understanding tasks. We have chosen to exclude table-to-text tasks from our study, since they involve open-ended questions without fixed or definitive answers. The detailed statistics for our training data are presented in Table 2, covering WTQ, HiTab, TabMWP, and TabFac. To assess the robustness of various optimization methods, we set aside similar tasks, such as TAT-QA, InfoTabs, as held-out. To evaluate performance, we employ the accuracy to evaluate overall reasoning performance and TEDs for perception evaluation.

Due to imbalanced dataset sizes, an equal number of entries from each are sampled. Our sampling process is carefully designed to create representative and manageable subsets for training and evaluation, particularly considering the computational constraints imposed by high-resolution images and long text sequences. Our procedure involves a three-step approach:

Image Resolution Filter: We exclude images with exceptionally large dimensions that could lead to out-of-memory errors. Specifically, any image with a total pixel count exceeding 1/8 of the Qwen2-VL model's maximum capacity (12,845,056 pixels) is removed.

Output Length Filter: For the PA-GRPO task, samples where the ground-truth structured texts (e.g., Markdown) exceed 2048 tokens are removed to prevent memory issues when generating very large tables.

Random Sampling: We proceed with the random sampling. First, we select a diverse set of table images from this filtered pool. Then, we sample questions corresponding to these chosen tables un-

Method	Resolution	Question Answering				Fact Verification		Avg. I.	Avg. O.
		TabMWP$_I$	WTQ$_I$	HiTab$_I$	TAT-QA$_O$	TabFact$_I$	InfoTabs$_O$		
Closed-Source LVLM									
OpenAI-o4-mini	UNK	86.70	78.20	44.80	56.16	84.70	78.30	73.60	**67.23**
GPT-4o	UNK	87.59	64.39	39.32	53.85	73.33	**79.50**	66.16	66.68
Gemini-2.5-Pro	UNK	**89.90**	**80.34**	**46.44**	56.29	**85.02**	77.15	**75.43**	66.72
Claude-3.5-Sonnet	UNK	83.30	71.80	41.31	**59.33**	60.08	70.30	64.12	64.82
Open-Source LVLM									
Qwen2-VL-2B	Dyn.	46.10	22.30	22.90	30.44	8.90	24.60	25.05	27.52
DeepSeek-VL2 4.5B	Dyn.	53.75	40.42	18.89	24.11	13.65	24.79	31.68	24.45
mPLUG-Owl2 7B	448	6.83	0.67	0.13	0.39	8.21	26.19	3.96	13.29
Monkey 7B	896	13.26	19.07	6.41	12.31	22.56	22.11	15.33	17.21
Qwen2-VL-7B-Ins	Dyn	49.51	19.73	5.33	20.85	40.00	46.56	28.64	33.71
Qwen2-VL-7B	Dyn.	63.80	46.50	33.10	46.50	7.40	32.80	37.70	39.65
LLaVA-v1.5 7B	336	6.05	1.24	2.03	2.97	18.9	28.31	7.06	15.64
Table-LLaVA 7B	336	57.78	18.43	10.09	12.82†	59.85	65.26†	36.54	39.04
Table-LLaVA 13B	336	59.77	20.41	10.85	15.67†	65.00	66.91†	39.01	41.29
Qwen2-VL-72B	Dyn.	81.95	65.70	44.04	52.23	73.45	72.82	66.29	62.53
QVQ-72B-Preview	Dyn.	**86.20**	**68.20**	**45.70**	**55.48**	**77.68**	**74.64**	**69.45**	**65.06**
Optimizated LVLM									
Qwen2-VL-2B-SFT	Dyn.	70.00	31.00	31.90	19.69	51.70	37.40	46.15	28.55
Qwen2-VL-2B-GRPO	Dyn.	71.40	**35.30**	35.20	**29.02**	22.70	29.00	41.15	29.01
Qwen2-VL-2B-Table-R1	Dyn.	**83.20**	34.40	**37.30**	26.42	**60.90**	**43.60**	**53.95**	**35.01**
Qwen2-VL-7B-SFT	Dyn.	90.30	46.80	48.50	37.82	73.20	57.60	64.70	47.71
Qwen2-VL-7B-GRPO	Dyn.	89.20	**53.20**	54.70	**51.68**	11.90	41.60	52.25	46.64
Qwen2-VL-7B-Table-R1	Dyn.	**92.60**	50.30	**58.20**	48.06	**73.40**	**62.80**	**68.63**	**55.43**

Table 3: Evaluation results on 4 held-in and 2 held-out multimodal tabular tasks. The subscripts $_I$ and $_O$ denote held-in and held-out, respectively, while † indicates the model has been trained on this dataset. "Dyn." denotes dynamic resolution processing, where input images are adaptively resized to preserve aspect ratios.

til our target of 8,000 entries is reached.

5.3 Experimental Setup

We strategically select two open-source LVLMs as our policy model: Qwen2-VL-2B and Qwen2-VL-7B (Wang et al., 2024b), since they have strong cognitive behaviors that enhance self-reflection on reasoning tasks (Gandhi et al., 2025). All experiments are conducted on 8 NVIDIA A100 80GB Tensor Core GPUs with DeepSpeed (Rajbhandari et al., 2020; Rasley et al., 2020), Zero stage 2, and HuggingFace Accelerate (Gugger et al., 2022). During the warm-up stage, we use AdamW optimizer (Loshchilov and Hutter, 2017) with a 10% warm-up ratio and 1000 steps. Following prior work (Chen et al., 2025), learning rates are set 2e^{-5} and 5e^{-6} respectively for Qwen2-VL-2B and Qwen2-VL-7B. Given the large image resolution shown in Table 2, we set batch sizes to 2 and 1.

For the PA-GRPO and HC-GRPO stage, we perform 4 rollouts per question ($G = 4$) and set the sampling temperature to 1 to encourage diverse reasoning trajectories. The maximum sequence length is set to $L = 1024$, ensuring that the model can generate complete reasoning paths. Both the policy model and reference model are initialized from the model after the warm-up, with the reference model frozen during training. The epoch and batch size are set to 2 and 1. Following (Chen et al., 2025), the KL divergence coefficient β in Eq. 7 is set to 0.04 by default, and the learning rate for the policy model is set to 1e^{-6} for both Qwen2-VL-2B and Qwen2-VL-7B.

5.4 Table Reasoning Performance

Table 3 depicts the comprehensive comparison of Table-R1 against baselines. By analyzing the experimental results, we have the following findings:

Open-source Model Hierarchy. Open-source models exhibit a clear performance hierarchy aligned with model size. Smaller models (Qwen2-VL-2B: Avg. I.=25.05%, Avg. O.=27.52%) significantly underperform their larger counterparts (QVQ-72B-Preview: Avg. I.=69.45%, Avg. O.=65.06%) by 44.40% and 37.54%, respectively. Notably, the 72B parameter class achieves performance comparable to closed-source models (GPT-4o: Avg. I=66.16%, Avg. O=66.68%), demonstrating the scalability of open architectures.

Closed-source Model Superiority. Closed-source models consistently outperform open-source models across most datasets. For example, Gemini-2.5-pro (Avg. I.=75.43%, Avg.

Dataset	Qwen2-VL-2B	Table-R1-2B	Table-LLaVA 7B
WTQ$_I$	0.41	**0.73**	0.56
TabMWP$_I$	0.48	**0.81**	0.80
TabFact$_I$	0.63	**0.93**	0.40
HiTab$_I$	0.24	**0.54**	0.32
InfoTabs$_O$	0.16	0.56	**0.74**
TAT-QA$_O$	0.47	**0.70**	0.57

Table 4: Table structure recognition performance (TEDS score) on four held-in (subscript I) and two held-out (subscript O) datasets. The best-performing score on each dataset is highlighted in **bold**. Table-R1-2B represents Qwen2-VL-2B-Table-R1.

Method	Question Answering		Fact Verification	
	Avg. QA$_I$	TAT-QA$_O$	TabFact$_I$	InfoTabs$_O$
Qwen2-VL-2B				
Table-R1	51.37	26.42	60.90	43.60
w/o Warm-up	38.87	23.16	20.50	26.10
△	-12.50	-3.26	-40.40	-17.50
w/o PA-GRPO	50.60	26.20	60.20	42.90
△	-0.77	-0.22	-0.70	-0.70
w/o HC-GRPO	36.27	20.08	45.80	41.90
△	-19.70	-6.34	-15.10	-1.70

Table 5: Effectiveness Across Different Stages. We report the average performance across various benchmarks, where Avg. QA_I denotes the average accuracy of three QA datasets. △ denotes the performance gap between Table-R1 and its variants.

	Number of HC splits per solution			
Dataset	1	2	3	4
TabMWP	81.5	82.6	83.2	83.4

	Number of generations G per question				
Dataset	2	3	4	5	6
TabMWP	81.2	82.5	83.2	83.4	83.5

Table 6: Impact of the number of splits per solution and generations per question on TabMWP's performance. Experiments are conducted on Qwen2-VL-2B.

O.=66.72%) outperforms the best open-source model by 5.98% in Avg. I. and 1.66% in Avg. O., suggesting stronger multimodal table reasoning abilities.

Optimization Methods Comparison. All optimized LVLMs demonstrate significant improvements over baseline models. Notably, Table-R1 achieves superior overall performance on both held-in and held-out. Specifically, for Qwen2-VL-7B, Table-R1 outperforms SFT by 3.93% and GRPO by 16.38% on held-in. On held-out, it surpasses SFT by 7.72% and GRPO by 8.79%. This performance notably exceeds that of Table-LLaVA 13B and can be comparable with GPT-4o on held-in.

GRPO Sensitivity to Initial Capability. While GRPO generally outperforms SFT across most QA tasks, a notable performance gap in table fact verification is observed. This discrepancy arises from the initial capabilities of the policy model. When it is significantly low (Qwen2-VL-7B: Avg. I.=7.40% on TabFact$_I$), the rewards derived from group responses tend to approach zero. This situation leads to low standard deviations in Eq. 6, which hinders the convergence of the reinforcement learning.

5.5 Table Perception Performance

As Table 4 shows, our Qwen2-VL-2B-Table-R1 demonstrates competitive or superior performance on most datasets compared to Table-LLaVA 7B. We note that Table-LLaVA's performance is higher on InfoTabs. This is expected, as Table-LLaVA was explicitly trained on the InfoTabs dataset, whereas for Table-R1, this was a held-out dataset. Our strong performance on held-out tasks like TAT-QA underscores the robustness of our approach.

5.6 Ablation Studies

Effects of the Number of G and HC Splits. We conduct the parameter analysis on the number of generations G in HC-GRPO with Qwen2-VL-2B over TabMWP$_I$, analyzing its impact on reasoning performance. Table 6 demonstrates that a larger G typically results in better performance, as the baseline reward is estimated as the average reward of all generated reasoning paths. A larger G leads to low variance and a more stable estimation of the baseline reward, making the optimization process more stable. However, increasing G also raises higher computational costs. Thus, $G = 4$ is set as the default to balance performance and computational efficiency. Similarly, increasing the number of HC splits enhances performance by generating more training data, with a default value of 3.

Effectiveness of Warm-up. Table 5 reveals that the impact of excluding the warm-up phase causes the most severe performance drop (e.g., -40.40% on TabFact$_I$), highlighting its necessity in mitigating GRPO's sensitivity to poor initial accuracy. Furthermore, integrating a warm-up stage consistently markedly improves accuracy rewards, as demonstrated by the steep initial increase in accuracy rewards in Figure 4. This enhancement can be attributed to the rapid acquisition of fundamental reasoning capability during the warm-up stage.

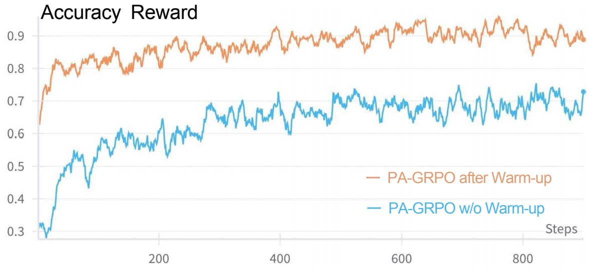


Figure 4: accuracy reward of PA-GRPO with and without warm-up.

Dataset	Qwen2-VL-2B-Table-R1	w/o PA-GRPO	Δ
WTQ$_I$	0.73	0.63	-0.10
TabMWP$_I$	0.81	0.53	-0.28
TabFact$_I$	0.93	0.85	-0.08
HiTab$_I$	0.54	0.41	-0.13
InfoTabs$_O$	0.56	0.53	-0.03
TAT-QA$_O$	0.70	0.62	-0.08

Table 7: Effectiveness of the PA-GRPO module on table perception performance of four held-in (subscript I) and two held-out (subscript O) datasets. 'w/o PA-GRPO' is the model variant without PA-GRPO, and Δ quantifies the resulting performance drop.

Methods	Warm-up(W)	W+PA-GRPO	W+SFT
TEDS	0.51	0.80	0.84
Accuracy	60.2	58.7	14.30

Table 8: Comparison of the warm-up, PA-GRPO, and SFT methods over TEDS and accuracy on TabMWP, based on training the Qwen2-VL-2B model.

Effectiveness of the PA-GRPO. Table 5 indicates that the absence of PA-GRPO yields only marginal performance changes. We conduct a comprehensive comparison of PA-GRPO and the standard SFT method on TabMWP, utilizing Qwen2-VL-2B as the foundational model. As shown in Table 8, both TEDS and accuracy are used to assess recognition and reasoning abilities, respectively. Although SFT improves the model's visual recognition (TEDS 0.84), it severely impairs reasoning accuracy from 60.2% to 14.30%. In contrast, PA-GRPO significantly boosts recognition performance while maintaining strong reasoning capabilities, demonstrating its effectiveness as a more balanced optimization strategy. More details on the effectiveness of the PA-GRPO module on table perception performance are shown in Table 7. Removing PA-GRPO causes a substantial drop in accuracy across all datasets (e.g., a 0.28 drop on TabMWP). This result directly confirms that PA-GRPO is highly effective at its intended task: improving the model's ability to accurately recognize and represent table structures.

Methods	$\mathbf{W_{qs}}$	$\mathbf{W_{qs}}$+GRPO	$\mathbf{W_{hc}}$	$\mathbf{W_{hc}}$+HC-GRPO
Accuracy	60.20	76.40	63.60	83.00

Table 9: Effectiveness of the different methods on TabMWP using Qwen2-VL-2B, where W_{qs} and W_{hc} denote warm-up stage with question-solution pairs and hint-completion pairs.

Effectiveness of the HC-GRPO. Our proposed HC-GRPO introduces a fine-grained, residual-step reward in contrast to conventional coarse solution-level rewards. To assess its effectiveness, we conduct a comparative study on TabMWP using Qwen2-VL-2B under four training configurations: (1) warm-up$_{qs}$ with question-solution pairs datasets; (2) warm-up$_{qs}$ + GRPO (solution-level reward); (3) warm-up$_{hc}$ using hint-completion pairs datasets; (4) warm-up$_{hc}$ + HC-GRPO (residual-step reward). As shown in Table 9, both warm-up$_{qs}$ and warm-up$_{hc}$ can improve performance. However, HC-GRPO can achieve higher accuracy. This reveals that residual-step rewards are more effective in enhancing LVLMs' reasoning capabilities, as they offer finer-grained supervision by aligning rewards with the remaining reasoning steps, thereby enabling more precise credit assignment than coarse solution-level rewards.

6 Conclusion

We introduce Table-R1, a novel three-stage framework that significantly enhances multimodal table perception and reasoning by integrating warm-up initialization, continuous reward refinement through PA-GRPO, and fine-grained hint-based reasoning with HC-GRPO. Through extensive evaluation, Table-R1 demonstrates superior performance and robustness compared to both SFT and GRPO methods. Additionally, it significantly outperforms existing open-source LVLM, even on par with the powerful GPT-4o on some benchmarks. Overall, our approach not only underscores the pivotal role of initial policy accuracy in reinforcement learning for reasoning tasks but also establishes a practical pathway for advancing RL-driven multimodal comprehension in real-world applications.

Limitations

Despite its promising performance, our framework faces three key limitations that motivate future research. First, the current framework focuses primarily on generating definitive answers, leaving significant room for exploration in the area of table text generation. For example, tasks such as table summarization and table description generation are not fully addressed. Second, our evaluation relies on English-only benchmarks with clear images, whereas real-world table images often exhibit perspective distortions, uneven lighting, or handwriting, and multilingual contexts remain unaddressed. Third, the HC-GRPO stage relies on coarse binary rewards for correctness and formatting; richer signals such as step-level validity scores or continuous semantic similarity metrics could yield more nuanced training.

Acknowledgments

We thank all anonymous reviewers for their valuable comments. The work was partially supported by the following: National Natural Science Foundation of China under No. 92370119, 62436009, 62276258 and 62376113, XJTLU Funding REF-22-01-002, Research Development Fund with No. RDF-22-01-020, and Suzhou Municipal Key Laboratory for Intelligent Virtual Engineering (SZS2022004). We would like to express our deepest gratitude to the Red Bird MPhil Program at the Hong Kong University of Science and Technology (Guangzhou) for providing us with generous support, resources, and funding.

Ethical Considerations

We discuss the following ethical considerations related to Table-R1: (1) **Intellectual Property.** We adhere to the license when using existing datasets, such as Apache-2.0 for MMTab and MIT for TabMWP. (2) **Intended Use.** Table-R1 can be utilized to develop more persuasive multimodal table reasoning models. Researchers can also inherit our methodological design to develop their RL models in other scenarios. (3) **Controlling Potential Risks.** Since the training of Table-R1 only includes public datasets, which do not require extensive judgments about social risks, we believe Table-R1 does not introduce any additional risks. We manually verified some randomly sampled data from the experimental datasets to ensure the dataset did not contain risky issues.

References

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1:1.

Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. 2025. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. https://github.com/Deep-Agent/R1-V. Accessed: 2025-02-02.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. *Preprint*, arXiv:1909.02164.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. Finqa: A dataset of numerical reasoning over financial data. *Proceedings of EMNLP 2021*.

Kanzhi Cheng, Yantao Li, Fangzhi Xu, Jianbing Zhang, Hao Zhou, and Yang Liu. 2024. Vision-Language Models Can Self-Improve Reasoning via Reflection. *Preprint*, arXiv:2411.00855.

Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. 2022. HiTab: A hierarchical table dataset for question answering and natural language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1094–1110, Dublin, Ireland. Association for Computational Linguistics.

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

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D. Goodman. 2025. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *Preprint*, arXiv:2503.01307.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate.

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

Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. 2020. INFOTABS: Inference on tables as semi-structured data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2309–2324, Online. Association for Computational Linguistics.

Xiaoqiang Kang, Zimu Wang, Xiaobo Jin, Wei Wang, Kaizhu Huang, and Qiufeng Wang. 2025. Template-driven llm-paraphrased framework for tabular math word problem generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24303–24311.

Yannis Katsis, Saneem Chemmengath, Vishwajeet Kumar, Samarth Bharadwaj, Mustafa Canim, Michael Glass, Alfio Gliozzo, Feifei Pan, Jaydeep Sen, Karthik Sankaranarayanan, and Soumen Chakrabarti. 2022. AIT-QA: Question answering dataset over complex tables in the airline industry. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 305–314, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

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

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024b. Monkey: Image resolution and text label are important things for large multi-modal models. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Yicheng Liu, Bowen Wang, and et al. 2025a. Understanding r1: On zero-like training and reward-conditioned policy optimization. *arXiv preprint arXiv:2503.11234*.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025b. Visual-RFT: Visual Reinforcement Fine-Tuning. *Preprint*, arXiv:2503.01785.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *International Conference on Learning Representations (ICLR)*.

Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. ReFT: Reasoning with Reinforced Fine-Tuning. *Preprint*, arXiv:2401.08967.

Suyash Vardhan Mathur, Jainit Sushil Bafna, Kunal Kartik, Harshita Khandelwal, Manish Shrivastava, Vivek Gupta, Mohit Bansal, and Dan Roth. 2024. Knowledge-aware reasoning over multimodal semi-structured tables. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14054–14073, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Panupong Pasupat and Percy Liang. 2015. 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)*, pages 1470–1480, Beijing, China. Association for Computational Linguistics.

Yi Peng, Chris, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyidan Xie, Li Ge, Rongxian Zhuang, Xuchen Song, Yang Liu, and Yahui Zhou. 2025. Skywork R1V: Pioneering Multimodal Reasoning with Chain-of-Thought. *Preprint*, arXiv:2504.05599.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of SIGKDD*.

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

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

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

Qwen Team. 2024. Qvq: To see the world with wisdom.

Boris Van Breugel and Mihaela Van Der Schaar. 2024. Position: Why tabular foundation models should be a research priority. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 48976–48993. PMLR.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. 2024a. Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations. *Preprint*, arXiv:2312.08935.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Shuhe Wang, Shengyu Zhang, Jie Zhang, Runyi Hu, Xiaoya Li, Tianwei Zhang, Jiwei Li, Fei Wu, Guoyin Wang, and Eduard Hovy. 2025. Reinforcement Learning Enhanced LLMs: A Survey. *Preprint*, arXiv:2412.10400.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

Bohao Yang, Yingji Zhang, Dong Liu, André Freitas, and Chenghua Lin. 2025. Does table source matter? benchmarking and improving multimodal scientific table understanding and reasoning. *Preprint*, arXiv:2501.13042.

Jingfeng Yang, Aditya Gupta, Shyam Upadhyay, Luheng He, Rahul Goel, and Shachi Paul. 2022. TableFormer: Robust Transformer Modeling for Table-Text Encoding. In *Proc. of ACL*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *Preprint*, arXiv:2311.04257.

Hang Yu, Tianle Li, and et al. 2025. Dapo: Data-aware policy optimization for open-source llm alignment. *arXiv preprint arXiv:2503.04111*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2025. Self-Rewarding Language Models. *Preprint*, arXiv:2401.10020.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024. ReST-MCTS*: LLM Self-Training via Process Reward Guided Tree Search. *Preprint*, arXiv:2406.03816.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025. R1-VL: Learning to Reason with Multimodal Large Language Models via Step-wise Group Relative Policy Optimization. *Preprint*, arXiv:2503.12937.

Weichao Zhao, Hao Feng, Qi Liu, Jingqun Tang, Binghong Wu, Lei Liao, Shu Wei, Yongjie Ye, Hao Liu, Wengang Zhou, Houqiang Li, and Can Huang. 2024. Tabpedia: Towards comprehensive visual table understanding with concept synergy. In *Advances in Neural Information Processing Systems*.

Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. 2024. Multimodal table understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9102–9124, Bangkok, Thailand. Association for Computational Linguistics.

Xinyi Zheng, Doug Burdick, Lucian Popa, Peter Zhong, and Nancy Xin Ru Wang. 2021. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. *Winter Conference for Applications in Computer Vision (WACV)*.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. 2020. Image-based table recognition: Data, model, and evaluation. In *Computer Vision – ECCV 2020*, pages 564–580, Cham. Springer International Publishing.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. 2019. Image-based table recognition: data, model, and evaluation. *arXiv preprint arXiv:1911.10683*.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online. Association for Computational Linguistics.

A Pseudocode of Table-R1

The overall training algorithm of Table-R1 is presented in Algorithm 1.

Algorithm 1 Pseudocode of our Table-R1

```
Input: Policy model $\pi_\theta$ initialized by a pre-trained LVLM; a vision-text dataset $D_p$ and $D_r$.
Output: Trained policy model $\pi_\theta$
Policy warm-up:
for iter = 1 to $N$ do
    Sample $\{I, Q, S\} \in D_s \cup D_p$
    Optimize policy model $\pi_\theta$ by SFT
end for
Perception Stage:
for iter = 1 to $N$ do
    Sample $\{I, Q_p, S_p\} \in D_p$
    Generate a group of perception paths $\{\mathbf{c}^i\}_{i=1}^M \sim \pi_\theta$
    Obtain Tree-Edit-Distance-based Similarity (TEDS) as rewards $\{r^i\}_{i=1}^M$
    Obtain relative advantages $\{\hat{A}^i\}_{i=1}^M$ by Eq. 6
    Optimize policy model $\pi_\theta$ by Eq 7
end for
Reasoning Stage:
for iter = 1 to $N$ do
    Sample $\{I, Q_r, S_r\} \in D_r$
    Generate a group of reasoning paths $\{\mathbf{c}^i\}_{i=1}^M \sim \pi_\theta$
    Obtain accuracy rewards and format rewards $\{r^i\}_{i=1}^M$
    Obtain relative advantages $\{\hat{A}^i\}_{i=1}^M$ by Eq. 6
    Optimize policy model $\pi_\theta$ by Eq 7
end for
return policy model $\pi_\theta$
```

B More Details about Table-R1

B.1 Prompts for GRPO Training

To ensure correct output formatting during the training of PA-GRPO and HC-GRPO, we adopt the prompts presented in Table 10.

Table 10: Prompts used in PA-GRPO and HC-GRPO.

Perception Prompt: A conversation between User and Assistant. The user asks a question, and the Assistant solves it. This task is a simple perception task, and the Assistant directly provides the answer within the <answer> </answer> tags. For example: <answer> answer here </answer>
Hint-Completion Prompt: A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think><answer> answer here without unit </answer>

B.2 Instruction variants for PA-GRPO Task

In the stage of PA-GRPO in Section 4.2, we generate various instructions to transform the image into structured content. All possible templates are listed in Figure 5. During the construction of the D_p, we randomly select one from them.

```
"Please read the table in this image and return
a markdown-style reconstructed table in text.",
"Take a look at the table in this image and
provide me with the markdown representation of
the table in text format.",
"Read the shown table in this image and give me
the reconstructed table in the markdown text
format.",
"Watch the table in this image and convert it
into a Markdown table in the text form.",
"Given a table image, can you convert the table
into a Markdown table in text form?",
"Reconstruct the table in this picture as a
markdown-style table in text.",
"Please review this table image and return a
text representation of the table in the markdown
 format.",
"Examine the table in the shown picture and
generate a markdown text representation of the
table.",
"Watch this table and show a markdown-style
reconstructed table in text.",
"This picture illustrates a table. Please
represent this table with the markdown format in
 text.",
"Recognize the table in the presented picture
and represent it in the markdown format.",
"Recognize the table in this picture and return
a markdown-style reconstructed table in text.",
"Can you interpret the table in this image and
return it as a markdown table in text?"
"Look at the table in this image and reconstruct
 it as a markdown table in text format."
"Identify the table in this image and provide
its markdown text representation."
"Please examine the table in this image and
return it as a markdown table in text format."
"Can you read the table in this image and give
me the markdown table in text?"
"Please look at the table in this image and
provide the markdown table in text format."
```

Figure 5: Instruction variants for constructing the table recognition task.

B.3 Prompts for Long-COT Data Generation

In the warm-up stage in Section 4.2, some original solutions are too short to be divided into two parts. Therefore, we use the prompt, shown in Figure 6, to expand short-COT into long-COT formats.

B.4 Qualitative Breakdown of Failure Cases

Our error analysis identified five main categories of failures. In descending order of frequency, they are as follows:

- **Basic Arithmetic Calculation Errors:** The model frequently makes fundamental mis-

```
Instruction
Your role is to serve as a step-by-step solution
 provider for mathematical exercises. You will
generate a detailed explanation of the solution
to a given mathematical question, using the
provided table or data. Your explanations should
 be clear, logical, and adhere to the following
guidelines:
1. **Provide a Detailed Step-by-Step Solution**:
 Generate a comprehensive, step-by-step guide
that explains each part of the solution process,
 including the reasoning behind each step.
2. **Maintain the Integrity of the Answer**:
Ensure that the final answer provided in the
explanation aligns with the one given in the
input without altering its fundamental
correctness.
Here is an example of how the input and output
should be structured:
<an Example of Demonstration>
```

Figure 6: Prompt for constructing long-COT solution.

takes in basic mathematical operations (e.g., addition, subtraction).

- **Failure to Parse Complex Data Formats:** The model struggles to correctly interpret complex tables, such as those containing merged cells, or tables that are exceptionally long or wide.

- **Misunderstanding of Boundary Conditions:** This involves the poor interpretation of qualifying phrases (e.g., "at least," "more than"), leading to incorrect filtering of data.

- **Core Concept Confusion:** In these cases, the model misunderstands a core mathematical concept required by the question, such as 'absolute value'.

- **Omission of Key Information:** The least common error, this happens when the model fails to process the entire user prompt, overlooking crucial sentences or data points.