Think with Structured Grounding: Perceptual Reinforcement Learning for Chart and Visual-Tabular Understanding

Changjiang Jiang[1] **Qiannian Zhao**[1] **Lei Xin**[1] **Jinxiang Xie**[2] **Preslav Nakov**[1] **Zhuohan Xie**[1]

[1]Mohamed bin Zayed University of Artificial Intelligence [2]Nanjing University

Abstract

Multimodal Large Language Models (MLLMs) capable of "thinking with images" often rely on external tools for fine-grained perception. However, this reliance introduces significant inference latency and fails to effectively resolve the spatial-structural gap—a fundamental challenge in text-dense and structurally relational visuals (e.g., charts and visual tables) where strict relative spatial arrangements bind textual elements. Without external tools, standard MLLMs struggle with such fine-grained visual reasoning tasks. To address these issues, we propose Think with Structured Grounding (TwSG), a novel fine-grained image perception framework designed to internalize complex images's tool-use capabilities within the model. TwSG distills the benefits of multi-step reasoning and micro-cropping into a single efficient forward pass during inference. Specifically, we use an MLLM to identify key regions guided by ground-truth answers, and then prompt a teacher model to generate high-quality visual question-answering (VQA) data. These fine-grained, region-based supervisory signals are subsequently distilled back into the full-image representation. Our training pipeline consists of two stages: (1) a cold-start supervised fine-tuning (SFT) phase using multi-turn data with focused area descriptions to foster complex reasoning and error recovery; and (2) a reinforcement fine-tuning (RFT) phase driven by a novel process reward mechanism, TL-GRPO, which encourages strategic reasoning. Extensive experiments across various MLLM architectures demonstrate that TwSG reduces inference latency while substantially improving accuracy and robustness, endowing models with native fine-grained region description and flexible reasoning capabilities.

1 Introduction

Charts and visual-tabular data, as the most prevalent mediums for structured data visualization, are ubiquitous in real-world core applications (Masry et al., 2025a). Recently, Multimodal Large Language Models (MLLMs) have demonstrated remarkable capabilities in open-ended visual question answering and understanding (Chen et al., 2025). However, existing methods for Chart Question Answering (ChartQA) (Masry et al., 2022; Xu et al., 2025a) often overlook the text-dense and structurally relational nature of these formats. Specifically, charts typically feature abundant textual elements bound by strict relative spatial arrangements, creating what we term the *Spatial-Structural Gap*. This challenge is equally prevalent in visual-tabular QA tasks (Lompo & Haraoui, 2025). Previous studies have introduced "thinking with images" paradigms (Xu et al., 2025b), empowering models to invoke external tools such as dynamic cropping. While these approaches mitigate the initial perception gap, they inadvertently introduce a prohibitive efficiency bottleneck: frequent tool calling and repetitive patch inputs lead to a visual token explosion, resulting in severe inference latency and error accumulation during multi-turn interactions (Wei et al., 2026). Even when an MLLM perceives fine-grained details, natural language remains inherently too ambiguous to precisely navigate dense spatial layouts. Consequently, during the Chain-of-Thought (CoT) process, the model frequently fails to consistently anchor

Changjiang Jiang, Qiannian Zhao, Lei Xin: This work was completed during an internship at MBZUAI.
✉ {thejiangcj,zhuohan.xie18}@gmail.com

its linguistic reasoning to specific visual coordinates, leading to logical collapse and "spatial hallucinations" (e.g., misattributing a value to the incorrect row or bar).

To address these issues, we propose TwSG, a structured visual data distillation and reasoning framework. Our core idea is to shift the "interleaved visual reasoning process," which originally relied on external tools during the inference stage, into parameter internalization during the training stage. Specifically, when invoking external high-precision teacher models and image scaling tools, we directly distill the complex multi-turn image tool invocation generated by the teacher model into a single forward pass of the student model. To further unlock the MLLM's CvTR ability potential, we introduce a Reinforcement Learning (RL) phase following the cold-start. However, we observe that previous methods typically compute importance sampling ratios at either the token or sequence level, failing to account for the extreme variance in length across different functional tags. Specifically, the dense text information within `think` tags often results in much longer sequences compared to the concise logic in `answer` tags. This disparity leads to a **length-induced gradient bias**, where the optimization process is dominated by perceptual segments, potentially diluting the gradient contribution of critical reasoning steps. In addition, we propose TL-GRPO, a specialized reinforcement learning algorithm for TwSG's reasoning pattern. We introduce **Tag-level Importance Sampling (TL-IS)**, which normalizes importance weights independently within each tag to ensure the optimization is invariant to length disparities. To further stabilize the training, we design **Clipped Group Sampling (CGS)** to filter statistical outliers in advantage estimation. This synergy, combined with an interleaved verifiable reward mechanism, provides fine-grained guidance for the complex reasoning process, encouraging the spontaneous emergence of more strategic multi-step reasoning capabilities.

Our main contributions are summarized as follows:

- We systematically identify the issues of current MLLMs in CvTR tasks. Specifically, we reveal the inherent flaws associated with the recent "think with images" paradigm that heavily relies on external tool invocations such as prohibitive inference latency, token explosion, and error accumulation.
- We propose **TwSG**, a novel paradigm that internalizes tool capabilities via data distillation and process reward-driven reinforcement learning. Extensive experiments demonstrate that our approach significantly enhances reasoning accuracy and robustness while drastically reducing inference latency compared to state-of-the-art (SOTA) methods in the domain.
- We introduce **TL-GRPO**, a specialized reinforcement learning algorithm tailored for structured thinking trajectories. By integrating *Tag-level Importance Sampling*, *Clipped Group Sampling*, and an *Interleaved Verifiable Reward* mechanism, TL-GRPO effectively mitigates length-induced gradients bias and stabilizes the optimization of complex, multi-turn reasoning paths.

2 Related Work

2.1 CvTR

The field of Chart and Visual-Tabular Reasoning (CvTR) aims to empower MLLMs with the ability to interpret and reason over data-intensive visual structures. Early research in this domain primarily focused on single-modality table understanding (Jiang et al., 2025) or text-based TableQA (Wang et al., 2024a). However, the emergence of benchmarks like ChartQA (Masry et al., 2022) and its more sophisticated extension, ChartQAPro (Masry et al., 2025a), has shifted the focus toward interpreting complex visual marks (e.g., bars, sectors) and performing multi-step arithmetic. Parallelly, TableVQA-Bench (Kim et al., 2024) transitioned traditional text-based tabular datasets into the visual-tabular domain, demanding models to possess both high-fidelity text Recognition and spatial logical reasoning capabilities.

General-purpose MLLMs, such as Qwen3-VL (Bai et al., 2025a) and MiniCPM-V (Yu et al., 2025b), have demonstrated remarkable zero-shot performance on basic chart tasks. To further push the boundaries, domain-specific models like Chart-R1 (Chen et al., 2025) and Chart-RVR (Sinha et al., 2025) have been introduced, achieving state-of-the-art (SOTA) results through specialized fine-tuning. Additionally, CodeVision (Guo et al., 2026) explored a tool-augmented approach by converting image operations and cropping into executable

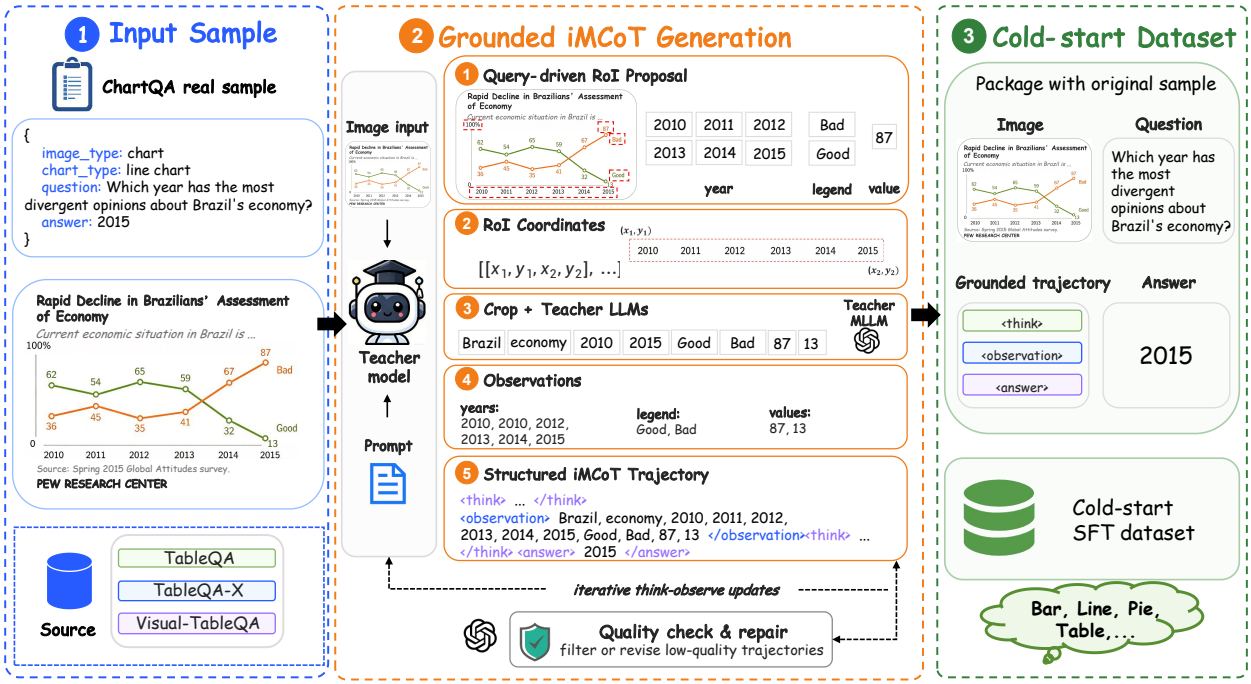

Figure 1: Pipeline for cold-start SFT data generation.

code within a Chain-of-Thought (CoT) framework. Despite these advances, existing methods often treat chart and tabular reasoning as isolated tasks. As observed in our experiments, models like Chart-RVR exhibit a performance trade-off, where optimization for charts leads to a degradation in visual-tabular understanding. Even MoE-based approaches like ChartMoE (Xu et al., 2025a), while effective for multi type chart scenarios, do not explicitly resolve this cross-format tension. Our work, TwSG, addresses this gap by seeking a unified reasoning perspective for both data formats, which are frequently co-present in critical domains like finance.

2.2 Interleaved Multimodel Chain-of-Thought

The concept of "Thinking with Images" has recently gained traction as a means to alleviate the limitations of MLLMs in perceiving fine-grained visual details. This paradigm typically involves the model autonomously calling image-cropping or enhancement tools to zoom into local regions, thereby improving grounding accuracy. For instance, in text perceptional tasks, VACoT (Xu et al., 2025b) proposed integrating image data augmentation tools into the reasoning process to resolve ambiguities in dense text recognition. Recent advancements have further diversified these visual interaction strategies: DeepEyes (Zheng et al., 2026) introduced the foundational approach of invoking specialized tools for adaptive image cropping, while DeepEyesV2 (Hong et al., 2026) evolved this into a more flexible framework by utilizing code execution to perform precise cropping and information retrieval. Furthermore, ThyME (Zhang et al., 2025) emphasized the necessity of temporal coherence in visual reasoning, employing multi-round cropping tool invocations to iteratively refine the model's perception.

However, a wide spectrum of prior efforts—including direct QA and forgery localization methods (e.g., IML (Qu et al., 2023), IML2 (Qu et al., 2024), RTM (Qu et al., 2025), Omni-IML (Qu et al., 2026a), DS-Net (Qu et al., 2026b), Mesorch (Zhu et al., 2025), Imdl-benco (Ma et al., 2024), Forensichub (Du et al., 2026), RIML (Zhu et al., 2026b), and Venus-DeFakerOne (Team, 2026)) as well as broader multimodal modeling frameworks (e.g., UniMoMo (Xin et al., 2026a), Beyond Human Annotation (Ying et al., 2026), IMG (Lin et al., 2025a), DualCPT (Xin et al., 2026b), Multi-Omics (Xin et al., 2024), TRS (Kong et al., 2025) and Hytrec (Xin et al., 2026c))—primarily investigate MLLM performance within standard, direct-QA or fixed-input paradigms. While these approaches have achieved notable empirical success on closed benchmarks, they heavily rely on monolithic representations without explicit intermediate visual verification. As a result,

their generalization ability proves inherently limited when transferred to structured, multi-hop reasoning tasks that demand fine-grained visual-symbolic alignment.

While iMCoT has shown promise in general visual grounding (Qi et al., 2026), its application to the CvTR domain remains underexplored. CvTR tasks are inherently "text-dense" and "computation-intensive", requiring both precise value extraction from small visual elements and high-level logical synthesis. Existing iMCoT frameworks have not yet been optimized for the structured, graph-like nature of charts and tables. In this paper, we bridge this gap by introducing a structured graph-based "Thinking with Images" strategy, enabling TwSG to perform more robust and interpretable reasoning on complex visual data structures. The complete training parameters setup can be see in Appendix B.

3 Methodology

In this section, we introduce **Think with Structured Grounding (TwSG)**, a novel framework designed to bolster the CvTR reasoning of MLLMs on complex document data through an internalized, self-consistent cognitive process.

3.1 Cold-start SFT Data Construction

The core philosophy of TwSG is to leverage a powerful teacher model to perform fine-grained, region-based perception and reasoning, and then condense this process into a student model that can perform "one-pass" reasoning without external tools. As shown in Figure 1, the TwSG pipeline consists of four main stages: (1) **Query-Driven Region Proposal:** Based on the given question and the global input image, a robust teacher model is employed to predict and return the most probable regions of interest (RoIs) that contain the target information; (2) **Sub-image Cropping and Teacher MLLMs Integration:** The retained RoIs are cropped into high-resolution sub-images to preserve fine-grained visual details. A teacher model is then invoked to extract highly accurate textual data from these specific crops; (3) **Structured iMCoT Trajectory Generation:** The teacher model outputs a detailed reasoning path along with the final answer. We summarize and format this output into a structured, iMCoT sequence. Crucially, to handle complex queries, this process accommodates multiple iterations of thinking and observing before reaching a conclusion. This encapsulates the internal cognitive reasoning and external perceptual steps using explicit tags, resulting in a multi-turn trajectory patterned as: `<think>` *[reasoning step 1]* `</think>` `<observation>` *[visual cue 1]* `</observation>` `<think>` *[reasoning step 2]* `</think>` `<observation>` *[visual cue 2]* `</observation>` ... `<answer>` *[final answer]* `</answer>`; (4) **Hallucination-Aware Refined Distillation:** To minimize hallucination in distilled trajectories, we utilize an independent teacher model as a Reward Model to audit the reasoning process. Instead, we instruct the teacher model to generate a corrective reasoning path. If an inconsistency is detected, rather than simply discarding the sample, we instruct the teacher model to generate a corrective reasoning path. Specifically, each "observation" is formatted as a structured JSON object, comprising bbox coordinates and the associated text fields, ensuring a precise mapping between visual grounding and textual evidence. Finally, we collect 12,674 Cold-Start SFT samples, denoted as TwSG-12K. Detailed data distributions and construction prompts are provided in Appendix E.

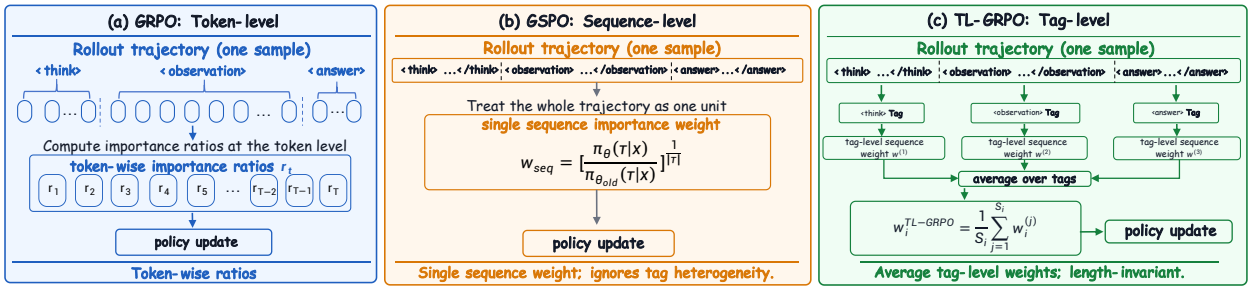

Figure 2: Comparison between Tag-level Importance Sampling and token-level or sequence-level importance sampling.

Unlike traditional CoT which provides a single reasoning block (Wei et al., 2026), we adopt an interleaved *Think-Observation-Answer* pattern (Xie et al., 2025). Specifically, for a query Q, the teacher generates a trajectory $\tau = \{t_1, o_1, t_2, o_2, \ldots, a\}$, where: (1) t_i (*Think*): The reasoning step focusing on a specific part of the structured graph G; (2) o_i (*Intermediate Evidence*): The extracted region-based perceptional result or sub-calculation result corresponding to t_i; (3) a (*Final Answer*): The final answer to the question. This format ensures that the model grounds each reasoning step in explicit visual evidence before proceeding to the next logical deduction.

3.2 TL-GRPO

Tag-level Importance Sampling. As shown in Figure 2, standard reinforcement learning frameworks, such as GRPO (Shao et al., 2024) and GSPO (Zheng et al., 2025), typically compute importance sampling ratios at either the granular token level or the sequence level. Although several recent works, such as EA-RLVR (Zhou et al., 2026), Se-GUI (Yuan et al., 2025), FakeVLM-R1 (Zhu et al., 2026a), EGPO (Zhao et al., 2026), Veritas (Tan et al., 2026a), and Veritas++ (Tan et al., 2026b), have modified multi-turn sampling schemes or reward functions, they still do not apply importance sampling from the visual Chain-of-Thought (CoT) perspective. Other approaches either adopt coarse-grained multi-turn rollout strategies, such as Fake-HR1 (Jiang et al., 2026b), or reformulate the overall GRPO objective along the visual CoT sequence, such as Ivy-Fake (Jiang et al., 2026a). Crucially, none of these methods account for the significant variance in sequence length across different functional tags. In contrast, our proposed sampling paradigm bridges the spatial-structural gap and substantially improves sampling efficiency and accuracy.

Specifically, during our cold-start phase, the model is trained to internalize dense region-based perceptual information while maintaining concise logical deductions, resulting in a highly non-uniform token distribution. Standard sampling techniques tend to be biased toward these longer perceptual segments, thereby diluting the gradient contributions of critical yet concise reasoning steps. To overcome this limitation, we propose **Tag-level Importance Sampling (TL-IS)**. By segmenting the trajectory according to our structured tripartite format, TL-IS normalizes the importance weight independently within each functional tag. This ensures that the optimization process remains invariant to length disparities between cognitive reasoning and perceptual observation, enabling more balanced and stable policy updates.

Following the above motivation, we represent a structured trajectory as a concatenation of tag-level sequences: $\tau = \{t_1, o_1, t_2, o_2, \ldots, a\}$, where each segment corresponds to a structured reasoning stage. We denote each tag segment as $\tau_i^{(j)}$ with length $|\tau_i^{(j)}|$.

We first compute **the sequence importance sampling ratio** within each tag segment:

$$w_i^{(j)} = \left[\frac{\pi_\theta(\tau_i^{(j)} \mid x)}{\pi_{\theta_{\text{old}}}(\tau_i^{(j)} \mid x)} \right]^{\frac{1}{|\tau_i^{(j)}|}} = \exp\left(\frac{1}{|\tau_i^{(j)}|} \sum_{t \in \tau_i^{(j)}} \log \frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid x, y_{i,<t})} \right). \tag{1}$$

We then aggregate across all tag segments to obtain the final tag-level importance weight:

$$w_i^{\text{TL-GRPO}} = \frac{1}{S_i} \sum_{j=1}^{S_i} w_i^{(j)}, \tag{2}$$

where S_i is the number of valid tag segments in τ_i.

Reward function. To guide the model's reasoning process and ensure output quality, we design a multi-dimensional reward function $R(y)$ consisting of three components: format integrity, answer accuracy, and region-based perceptional grounding. Formally, the total reward is defined as:

$$R(y) = \lambda_1 r_{\text{fmt}} + \lambda_2 r_{\text{ans}} + \lambda_3 r_{\text{verify}} \tag{3}$$

where: (1) **Format Reward (r_{fmt}):** We define r_{fmt} as an indicator function, where $\mathcal{F}$ is the set of sequences that strictly adhere to the template $\langle\text{think}\rangle\langle\text{observation}\rangle\langle\text{answer}\rangle$ without any additional or unauthorized

tags; (2) **Answer Reward (r_{ans}):** This term evaluates the semantic similarity between the predicted answer y_{ans} and the ground truth $\hat{y}$ using ANLS, i.e., $r_{\text{ans}} = \text{ANLS}(y_{\text{ans}}, \hat{y})$; (3) **Verify Reward ($r_{\text{verify}}$):** To ensure the model remains grounded in the visual evidence, we leverage an expert MLLM to evaluate the correctness of the intermediate observation $y_{\text{observation}}$ given the image I. Specifically, follow by OpenAI's practice (OpenAI, 2024), we formulate observation verification as a multiple-choice evaluation task with five options, A-E. For each intermediate observation, the judge MLLM assigns a positive score only when selecting Option A or B, which indicates that the observation is sufficiently accurate and visually grounded. All other options are treated as incorrect or hallucinated observations and receive a score of 0. Formally, given a trajectory containing N observation fields, the verification reward is computed as

$$r_{\text{verify}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(a_i \in \{A, B\}\right), \tag{4}$$

where a_i denotes the judge's selected option for the i-th observation. This normalization constrains r_{verify} to $[0, 1]$ regardless of the number of generated observations.

This sparse yet reliable reward penalizes ungrounded observations and suppresses hallucination propagation, while providing process-level supervision for generating high-quality reasoning traces. During training, we use Qwen3-VL-72B-Instruct (Bai et al., 2025a) as the judge MLLM to provide verification signals for process-level reward estimation. The detailed judgment prompt is provided in Appendix D.

Clipped Group Sampling. To further stabilize the reinforcement learning process, especially for tasks with high variance such as OCR and mathematical reasoning, we introduce **Clipped Group Sampling (CGS)**. Inspired by the dynamic sampling in DAPO (Yu et al., 2025a) and advantage filtering in CPPO (Lin et al., 2025b), CGS refines the advantage distribution by trimming statistical outliers within each group. Specifically, for a group of n sampled trajectories, we first compute their raw relative advantages $\hat{A}_i$. We then exclude the k trajectories with the highest and lowest advantage scores (typically $k = 1$), retaining the central $n - 2k$ trajectories for the policy update. This ensures that the advantage distribution remains centered and symmetric, preventing the model from over-fitting to anomalous lucky successes or being distracted by singular catastrophic failures. For the data used in TL-GRPO, following DeepSeek-R1 (Guo et al., 2025), we applied rejection sampling using the post-SFT checkpoint across the original SFT dataset and the training sets of ChartQA (Masry et al., 2022), ChartQA-X (Hegde et al., 2025), and Visual-TableQA (Lompo & Haraoui, 2025). Specifically, we generated responses four times per sample, discarding instances that were correctly answered in all attempts and retaining only those with at least one incorrect response. Ultimately, this yielded a final dataset of 64,334 entries for RFT. By synergizing these two mechanisms, our framework achieves a more robust credit assignment, effectively bridging the gap between fine-grained image perception and high-level logical deduction.

4 Experiment

4.1 Experiment Setup

Benchmarks. We evaluate the performance of TwSG across diverse CvTR tasks, we conduct experiments on three benchmarks: (1) TableVQA-Bench Kim et al. (2024): This benchmark focuses on open-domain visual tabular reasoning over complex tabular data; (2) ChartQA Masry et al. (2022): A large-scale benchmark designed for question answering about charts that involves both visual and logical reasoning. It combines human-written and machine-generated questions, necessitating multi-step arithmetic operations and the ability to link visual marks (e.g., bars, lines) to their corresponding data values; (3) ChartQAPro Masry et al. (2025a): As a more challenging extension of ChartQA, ChartQAPro incorporates a higher degree of visual and topical diversity, including real-world infographics and complex dashboards. Additionally, we compare our performance with closed-source models on CharXiv-R Wang et al. (2024b) in Appendix C.

Baselines. We evaluate MLLMs on these representative categories: (1) General MLLMs, including Qwen3-VL-Instruct Bai et al. (2025a), Qwen2.5-VL Bai et al. (2025b), MiniCPM-V-4.5 Yu et al. (2025b), and

Table 1: Performance comparison on chart and visual-tabular reasoning benchmarks. Accuracy (%) is reported for the machine-generated (M) and human-generated (H) subsets of ChartQA, five question types of ChartQAPro, and four data sources of TableVQA-Bench. Avg. denotes the macro-average over all 11 subsets.

Method	Size	Chart							Visual Tabular				Avg.
		ChartQA		ChartQAPro					TableVQA-Bench				
		M	H	Factoid	MCQ	Convers.	FactChk.	Hypoth.	VWTQ	VWTQ-Syn	VTabFact	FinTabNetQA	
General MLLMs													
LLaVA-OneVision-1.5	4B	95.44	78.80	37.34	43.46	31.48	52.46	43.86	60.11	65.55	83.20	77.58	60.84
	8B	94.64	78.88	38.88	35.05	34.72	53.28	46.38	62.20	71.33	84.00	76.82	61.47
Qwen2.5-VL	3B	94.24	74.24	29.24	38.32	33.86	49.59	38.91	56.27	65.59	75.60	77.76	57.60
	7B	94.88	80.56	38.88	49.53	34.66	56.56	37.73	61.67	70.61	80.00	72.88	61.63
Qwen3-VL	4B	94.24	73.28	29.24	38.32	33.86	49.59	38.91	56.17	61.77	76.80	73.19	56.85
	8B	94.72	76.40	39.53	55.14	36.81	60.25	39.83	59.56	63.29	80.80	77.05	62.13
MINICPM-V-4.5	8.7B	81.60	67.44	**47.39**	48.13	41.16	56.56	39.54	71.48	76.85	89.60	78.80	63.50
Reasoning MLLMs													
Visionary-R1	3B	92.40	69.84	27.41	31.31	31.57	36.48	36.58	50.63	62.10	74.80	76.33	53.59
M2-Reasoning	7B	88.80	80.32	37.68	54.67	41.93	53.28	55.90	72.51	78.65	90.00	76.43	66.38
VL-Rethinker	7B	83.12	72.72	43.36	60.75	41.27	60.66	53.78	68.41	70.79	86.00	72.68	64.87
CvTR-domain MLLMs													
Chart-RVR	3B	93.76	75.76	30.07	51.40	32.58	49.59	37.05	54.37	63.91	81.20	68.10	57.98
Chart-RVR-Hard	3B	92.88	80.08	30.88	54.67	36.32	50.00	40.32	54.97	62.89	83.60	70.16	59.71
Chart-R1	7B	95.20	88.72	40.63	52.80	40.62	63.93	51.71	71.39	77.98	87.60	78.74	68.12
TwSG	4B	95.36	86.48	37.88	58.92	40.73	61.02	50.91	72.83	79.66	88.40	77.52	68.16
	8B	**96.12**	**89.44**	45.92	**62.35**	**44.80**	**66.18**	**58.77**	**79.93**	**88.72**	**95.60**	**82.92**	**73.70**

LLaVA-OneVision-1.5 An et al. (2025); (2) Reasoning MLLMs, including M2-Reasoning AI et al. (2025), VL-Rethinker Wang et al. (2025), and Visionary-R1 Xia et al. (2025); (3) Chart and Visual-Tabular (CvTR)-domain MLLMs, including Chart-R1 Chen et al. (2025) and Chart-RVR Sinha et al. (2025). TwSG's backbone is Qwen3-VL-8B.

4.2 Main Result

As shown in Table 1, we compare MLLMs at comparable model scales on a unified suite of chart and visual-tabular reasoning (CvTR) benchmarks. TwSG-8B achieves the best overall performance, with an average accuracy of **73.70%**, outperforming the strongest prior model in this comparison, Chart-R1, by 68.12%. Moreover, TwSG-8B obtains the best performance on 10 out of the 11 evaluated subsets, demonstrating consistently strong generalization across both chart and visual-tabular reasoning tasks.

Strong performance on standard chart benchmarks does not necessarily translate into robust performance across more challenging chart and visual-tabular reasoning tasks. General-purpose MLLMs often achieve strong accuracy on ChartQA, yet their performance does not consistently transfer to the more reasoning-intensive ChartQAPro and TableVQA-Bench subsets. In contrast, TwSG exhibits more balanced performance across these benchmarks. Notably, even TwSG-4B achieves an average accuracy of **68.16%**, slightly surpassing the 7B Chart-R1 model (68.12%) despite using a smaller model scale. This further demonstrates the effectiveness of unified chart and visual-tabular reasoning.

For example, compared with its base model Qwen2.5-VL-3B, Chart-RVR-Hard improves ChartQA-H from 74.24% to 80.08%, while its FinTabNetQA accuracy decreases from 77.76% to 70.16%. These results suggest that improvements from chart-specific specialization do not necessarily transfer to structured visual-tabular data, motivating a unified treatment of the two modalities.

Figure 3 compares the inference throughput of different reasoning paradigms. Tool-augmented approaches, particularly the Think-with-Images series, incur substantial inference overhead due to external tool invocation and repeated perception–reasoning interactions. In contrast, our approach internalizes bounding-box reasoning and spatial grounding into the model's native reasoning process, avoiding external tool calls during inference. TwSG therefore maintains approximately 1.75–2.00 samples per second and achieves higher throughput than the Qwen3-VL base model, while substantially improving CvTR performance. Together with the accuracy results in Table 1, these results demonstrate a favorable accuracy–efficiency trade-off through unified, tool-free end-to-end reasoning.

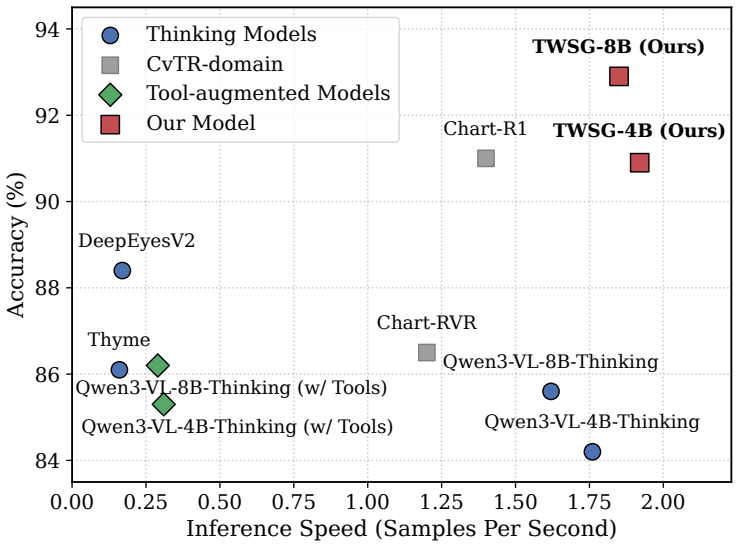

Figure 3: Samples per second.

4.3 Ablation Study

Table 2: Ablation studies of TL-GRPO on ChartQAPro and TableVQA-Bench. *w/o R* denotes the SFT checkpoint before reinforcement learning. Best results are shown in bold.

Effect of TL-GRPO				
Benchmark	GRPO	GRPO + TL-IS	GRPO + CGS	TL-GRPO (Full)
ChartQAPro	49.40	52.34	53.65	**55.60**
TableVQA-Bench	77.80	83.33	82.53	**86.79**
Reward Component Ablation				
Benchmark	w/o R	R_{ans}	$R_{\text{ans}} + R_{\text{fmt}}$	$R_{\text{ans}} + R_{\text{fmt}} + R_{\text{verify}}$
ChartQAPro	48.22	47.30	52.30	**55.60**
TableVQA-Bench	78.53	60.22	82.91	**86.79**
Reasoning Component Removal				
Benchmark	⟨`think`⟩	⟨`observation`⟩	⟨`answer`⟩	Full
ChartQAPro	48.23	51.72	54.32	**55.60**
TableVQA-Bench	80.92	79.95	85.50	**86.79**
CGS Hyperparameter K				
Benchmark	$K = 0$	$K = 1$	$K = 2$	$K = 3$
ChartQAPro	52.34	**55.60**	53.66	50.02
TableVQA-Bench	83.33	**86.79**	82.12	79.88

We conduct comprehensive ablation studies on ChartQAPro and TableVQA-Bench to validate the key design choices of our framework, with detailed results summarized in Table 2.

Comparison of TL-GRPO and Baseline. Both TL-IS and CGS consistently yield performance gains over the standard GRPO baseline. When combined, TL-GRPO achieves superior performance, outperforming GRPO by +6.20% on ChartQAPro (from 49.40% to 55.60%) and +8.99% on TableVQA-Bench (from 77.80% to 86.79%). This gain stems from a critical domain characteristic in visual reasoning: reasoning trajectories (⟨`think`⟩) are disproportionately long relative to the final prediction (⟨`answer`⟩). Under standard token-level

weighting, long reasoning steps dilute the gradient signal of compact answers, which tag-level sequence importance sampling effectively mitigates.

Effect of Cold Start SFT. The SFT checkpoint without RL (*w/o* R) achieves 48.22% on ChartQAPro and 78.53% on TableVQA-Bench. While cold start establishes fundamental instruction-following and question-answering capabilities, its generalization on complex visual-tabular reasoning remains limited without reinforcement learning.

Reward Function Decomposition. Evaluating individual reward terms reveals that all components are essential. Supervising solely with outcome correctness (R_{ans}) leads to severe performance degradation across both benchmarks (dropping to 47.30% and 60.22%), indicating optimization instability under sparse and noisy answer-only rewards. Incorporating formatting constraints (R_{fmt}) substantially restores and improves accuracy to 52.30% and 82.91%. Further adding visual verification (R_{verify}) achieves the best performance (55.60% / 86.79%), yielding additional gains of +3.30% and +3.88% respectively, which confirms the importance of fine-grained verification for structured reasoning.

Sensitivity of Clipping Hyperparameter K. Setting $K = 0$ corresponds to training with TL-IS alone without advantage-based trajectory clipping (achieving 52.34% and 83.33%). The optimal trade-off is achieved at $K = 1$ (55.60% / 86.79%). As K increases to 2 and 3, performance consistently degrades, because filtering out too many extreme-advantage samples excessively suppresses gradient variance and slows policy optimization.

Contribution of Reasoning Tags. Dissecting intermediate reasoning tokens shows distinct dependencies across tasks. Removing the ⟨`think`⟩ tag incurs the largest performance drop on ChartQAPro (-7.37%), while removing the ⟨`observation`⟩ tag causes the sharpest drop on TableVQA-Bench (-6.84%), verifying that explicit visual grounding is indispensable for structured tabular data. In contrast, removing ⟨`answer`⟩ results in a modest yet consistent degradation (-1.28% / -1.29%), demonstrating the utility of explicit answer boundary tokens in stabilizing output decoding.

5 Conclusion

In this paper, we identify that relying on the "think with images" paradigm for perceptual understanding in highly structured, text-intensive visual contexts introduces prohibitive inference latency and exacerbates the Spatial-Structural Gap in CvTR tasks. To overcome these bottlenecks, we propose TwSG, a novel structured data distillation framework tailored specifically for charts and tables. This framework significantly enhances the fine-grained OCR reasoning and perceptual capabilities of MLLMs. Furthermore, by introducing TL-GRPO and integrating a two-stage training paradigm that combines cold-start supervised fine-tuning with RL, we systematically align and bolster the model's performance in complex CvTR tasks. Extensive ablation studies and inference latency evaluations empirically demonstrate the effectiveness and efficiency of our proposed method. Ultimately, our work provides a highly efficient, tool-free pathway for advancing MLLMs in structurally complex visual reasoning.

Broader Impact Statement

Due to computational resource constraints, our framework was exclusively trained and evaluated on models with 8B parameters or fewer. Additionally, our approach inherently relies on the accuracy of the raw Cold start data. Consequently, the initial context fed into the large model cannot be guaranteed to be entirely error-free. Nevertheless, our empirical results demonstrate that our proposed method can still substantially enhance the CvTR capabilities of MLLMs despite this dependency.

Furthermore, we observe that computational reasoning challenges within chart understanding remain unaddressed, as pure CoT prompting falls short in handling complex arithmetic calculations. In contrast, TabDSR (Jiang et al., 2025) demonstrates the superiority of Program-of-Thought (PoT) in enhancing compu-

tational capabilities. Inspired by CodeVision (Guo et al., 2026), we plan to explore internalized, code-grounded image operations and programmatic visual reasoning in future work.

Our work aims to improve structured visual reasoning for charts and visual tables, which can benefit data analysis, document understanding, and accessibility-oriented applications. We do not develop technologies for weapons, biometric identification, surveillance, or direct decision-making in high-stakes domains. Therefore, we do not anticipate direct safety risks such as physical harm or increased weapon lethality.

References

Inclusion AI, :, Fudong Wang, Jiajia Liu, Jingdong Chen, Jun Zhou, Kaixiang Ji, Lixiang Ru, Qingpei Guo, Ruobing Zheng, Tianqi Li, Yi Yuan, Yifan Mao, Yuting Xiao, and Ziping Ma. M2-reasoning: Empowering mllms with unified general and spatial reasoning, 2025. URL https://arxiv.org/abs/2507.08306.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. Llava-onevision-1.5: Fully open framework for democratized multimodal training, 2025. URL https://arxiv.org/abs/2509.23661.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Lei Chen, Xuanle Zhao, Zhixiong Zeng, Jing Huang, Yufeng Zhong, and Lin Ma. Chart-r1: Chain-of-thought supervision and reinforcement for advanced chart reasoner, 2025. URL https://arxiv.org/abs/2507.15509.

Bo Du, Xuekang Zhu, Xiaochen Ma, Chenfan Qu, Kaiwen Feng, Zhe Yang, Chi-Man Pun, Ji-Zhe Zhou, et al. Forensichub: A unified benchmark & codebase for all-domain fake image detection and localization. *Advances in neural information processing systems*, 38, 2026.

Yuhan Fu, Ruobing Xie, Xingwu Sun, Zhanhui Kang, and Xirong Li. Mitigating hallucination in multimodal large language model via hallucination-targeted direct preference optimization. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16563–16577, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.850. URL https://aclanthology.org/2025.findings-acl.850/.

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

Zirun Guo, Minjie Hong, Feng Zhang, Kai Jia, and Tao Jin. Thinking with programming vision: Towards a unified view for thinking with images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33467–33476, 2026.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*, 2023.

Shamanthak Hegde, Pooyan Fazli, and Hasti Seifi. Chartqa-x: Generating explanations for visual chart reasoning, 2025. URL https://arxiv.org/abs/2504.13275.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. Deepeyesv2: Toward agentic multimodal model. In *ICLR*, 2026.

Changjiang Jiang, Fengchang Yu, Haihua Chen, Wei Lu, and Jin Zeng. Tabdsr: Decompose, sanitize, and reason for complex numerical reasoning in tabular data. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 3172–3196, 2025. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.169.

Changjiang Jiang, Wenhui Dong, Zhonghao Zhang, Fengchang Yu, Wei Peng, Xinbin Yuan, Yifei Bi, Ming Zhao, Zian Zhou, Chenyang Si, and Caifeng Shan. Ivy-fake: A unified explainable framework and benchmark for image and video aigc detection. In *Proceedings of the 2026 International Conference on Multimedia Retrieval*, ICMR '26, pp. 2438–2447. Association for Computing Machinery, 2026a. ISBN 9798400726170. doi: 10.1145/3805622.3810615. URL https://doi.org/10.1145/3805622.3810615.

Changjiang Jiang, Xinkuan Sha, Fengchang Yu, Jingjing Liu, Jian Liu, Mingqi Fang, Chenfeng Zhang, and Wei Lu. Fake-hr1: Rethinking reasoning of vision language model for synthetic image detection. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10482–10486, 2026b. doi: 10.1109/ICASSP55912.2026.11462736.

Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. Tablevqa-bench: A visual question answering benchmark on multiple table domains, 2024. URL https://arxiv.org/abs/2404.19205.

Zhenglun Kong, Yize Li, Fanhu Zeng, Lei Xin, Shvat Messica, Xue Lin, Pu Zhao, Manolis Kellis, Hao Tang, and Marinka Zitnik. Token reduction should go beyond efficiency in generative models–from vision, language to multimodality. *arXiv preprint arXiv:2505.18227*, 2025.

Junan Lin, Daizong Liu, Xianke Chen, Xiaoye Qu, Xun Yang, Jixiang Zhu, Sanyuan Zhang, and Jianfeng Dong. Audio does matter: Importance-aware multi-granularity fusion for video moment retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 6027–6036, 2025a.

ZhiHang Lin, Mingbao Lin, Yuan Xie, and Rongrong Ji. Cppo: Accelerating the training of group relative policy optimization-based reasoning models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Boammani Aser Lompo and Marc Haraoui. Visual-tableQA: Open-domain benchmark for reasoning over table images. In *NeurIPS 2025 Workshop on Foundations of Reasoning in Language Models*, 2025. URL https://openreview.net/forum?id=fvJRsGwhPf.

Xiaochen Ma, Xuekang Zhu, Lei Su, Bo Du, Zhuohang Jiang, Bingkui Tong, Zeyu Lei, Xinyu Yang, Chi-Man Pun, Jiancheng Lv, et al. Imdl-benco: A comprehensive benchmark and codebase for image manipulation detection & localization. *Advances in Neural Information Processing Systems*, 37:134591–134613, 2024.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. ChartQAPro: A more diverse and challenging benchmark for chart question answering. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and

Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19123–19151, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.978. URL https://aclanthology.org/2025.findings-acl.978/.

Ahmed Masry, Abhay Puri, Masoud Hashemi, Juan A. Rodriguez, Megh Thakkar, Khyati Mahajan, Vikas Yadav, Sathwik Tejaswi Madhusudhan, Alexandre Piché, Dzmitry Bahdanau, Christopher Pal, David Vazquez, Enamul Hoque, Perouz Taslakian, Sai Rajeswar, and Spandana Gella. Bigcharts-r1: Enhanced chart reasoning with visual reinforcement finetuning. In *Second Conference on Language Modeling*, 2025b. URL https://openreview.net/forum?id=19fydz1QnW.

Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. Chartassistant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7775–7803, 2024.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning, 2025. URL https://arxiv.org/abs/2503.07365.

OpenAI. Custom llm as a judge to detect hallucinations with braintrust. https://developers.openai.com/cookbook/examples/custom-llm-as-a-judge/, 2024. OpenAI Cookbook, accessed May 2026.

OpenRouter. Openrouter models. https://openrouter.ai/models?input_modalities=image, 2026. Accessed May 2026.

Yukun Qi, Pei Fu, Hang Li, Yuhan Liu, Chao Jiang, Bin Qin, Zhenbo Luo, and Jian Luan. Patchcue: Enhancing vision-language model reasoning with patch-based visual cues. *arXiv preprint arXiv:2603.05869*, 2026.

Chenfan Qu, Chongyu Liu, Yuliang Liu, Xinhong Chen, Dezhi Peng, Fengjun Guo, and Lianwen Jin. Towards robust tampered text detection in document image: New dataset and new solution. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5937–5946. IEEE, 2023.

Chenfan Qu, Yiwu Zhong, Chongyu Liu, Guitao Xu, Dezhi Peng, Fengjun Guo, and Lianwen Jin. Towards modern image manipulation localization: A large-scale dataset and novel methods. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10781–10790. IEEE, 2024.

Chenfan Qu, Yiwu Zhong, Fengjun Guo, and Lianwen Jin. Revisiting tampered scene text detection in the era of generative ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 694–702, 2025.

Chenfan Qu, Yiwu Zhong, Fengjun Guo, and Lianwen Jin. Omni-iml: Towards unified interpretable image manipulation localization. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Chenfan Qu, Yiwu Zhong, Xuekang Zhu, Junchi Li, Changjiang Jiang, Lianwen Jin, et al. Detect any ai-counterfeited text image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35437–35450, 2026b.

Qwen Team. Qwen3.6-Plus: Towards real world agents, April 2026. URL https://qwen.ai/blog?id=qwen3.6.

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

Sanchit Sinha, Oana Frunza, Kashif Rasul, Yuriy Nevmyvaka, and Aidong Zhang. Chart-rvr: Reinforcement learning with verifiable rewards for explainable chart reasoning, 2025. URL https://arxiv.org/abs/2510.10973.

Hao Tan, Jun Lan, Zichang Tan, Ajian Liu, Chuanbiao Song, Senyuan Shi, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. Veritas: Generalizable deepfake detection via pattern-aware reasoning. In *International Conference on Learning Representations*, 2026a.

Hao Tan, Jun Lan, Zichang Tan, Ajian Liu, Zijian Yu, Chuanbiao Song, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. Veritas++: Value-aware on-policy distillation for perception-enhanced aigi detection. *arXiv preprint arXiv:2607.27113*, 2026b.

GuangJian Team. Venus-defakerone: Unified fake image detection & localization. *arXiv preprint arXiv:2605.14091*, 2026.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning, 2025. URL https://arxiv.org/abs/2504.08837.

Jiacong Wang, Zijian Kang, Haochen Wang, LiangXiao, Ya Wang, Jiawen Li, Bohong Wu, Ran Jiao, Haiyong Jiang, ChaoFeng, and Jun Xiao. VGR: Visual grounded reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=kDhAiaGzrn.

Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. Chain-of-table: Evolving tables in the reasoning chain for table understanding. *ICLR*, 2024a.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2024b.

Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, et al. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. *arXiv preprint arXiv:2602.11858*, 2026.

Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning, 2025. URL https://arxiv.org/abs/2505.14677.

Roy Xie, David Qiu, Deepak Gopinath, Dong Lin, Yanchao Sun, Chong Wang, Saloni Potdar, and Bhuwan Dhingra. Interleaved reasoning for large language models via reinforcement learning. *arXiv preprint arXiv:2505.19640*, 2025.

Lei Xin, Caiyun Huang, Hao Li, Shihong Huang, Yuling Feng, Zhenglun Kong, Zicheng Liu, Siyuan Li, Chang Yu, Fei Shen, et al. Artificial intelligence for central dogma-centric multi-omics: Challenges and breakthroughs. *arXiv preprint arXiv:2412.12668*, 2024.

Lei Xin, Bin Gu, Peize Li, Zitong Wang, Jianbo Zhao, Changjiang Jiang, Yanyue Xie, Chao Huang, Xuyang Zhao, Zunhai Su, et al. Unimomo: Expert merging-based moe acceleration for large recommendation models. *arXiv preprint arXiv:2608.08627*, 2026a.

Lei Xin, Zhenglun Kong, Fukang Chen, Yuhao Zheng, Zeheng Wang, and Hao Tang. Dualcpt: Dual-branch modeling for cellular phenotype transition, 2026b.

Lei Xin, Yuhao Zheng, Ke Cheng, Changjiang Jiang, Zifan Zhang, and Fanhu Zeng. Hytrec: A hybrid temporal-aware attention architecture for long behavior sequential recommendation. *arXiv preprint arXiv:2602.18283*, 2026c.

Zhengzhuo Xu, Bowen Qu, Yiyan Qi, SiNan Du, Chengjin Xu, Chun Yuan, and Jian Guo. Chartmoe: Mixture of diversely aligned expert connector for chart understanding. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=o5TsWTUSeF.

Zhengzhuo Xu, Chong Sun, SiNan Du, Chen Li, Jing Lyu, and Chun Yuan. Vacot: Rethinking visual data augmentation with vlms, 2025b. URL https://arxiv.org/abs/2512.02361.

Dehao Ying, Fengchang Yu, Haihua Chen, Changjiang Jiang, Yurong Li, and Wei Lu. Beyond human annotation: Recent advances in data generation methods for document intelligence. *arXiv preprint arXiv:2601.12318*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025a.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe, 2025b. URL https://arxiv.org/abs/2509.18154.

Xinbin Yuan, Jian Zhang, Kaixin Li, Zhuoxuan Cai, Lujian Yao, Jie Chen, Enguang Wang, Qibin Hou, Jinwei Chen, Peng-Tao Jiang, and Bo Li. Se-gui: Enhancing visual grounding for gui agents via self-evolutionary reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 127658–127679. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/b95c7e24501f5d1dddbc5e8526cda7ae-Paper-Conference.pdf.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan Fan, Kaibing Chen, Jiankang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. Thyme: Think beyond images, 2025. URL https://arxiv.org/abs/2508.11630.

Qiannian Zhao, Chen Yang, Jinhao Jing, Yunke Zhang, Xuhui Ren, Lu Yu, Shijie Zhang, and Hongzhi Yin. Know what you know: Metacognitive entropy calibration for verifiable rl reasoning. *arXiv preprint arXiv:2602.22751*, 2026.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization, 2025. URL https://arxiv.org/abs/2507.18071.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In *ICLR*, 2026.

Jiang Zhou, Xiaohu Zhao, Xinwei Wu, Tianyu Dong, Hao Wang, Yangyang Liu, Heng Liu, Linlong Xu, Longyue Wang, Weihua Luo, and Deyi Xiong. Incentivizing parametric knowledge via reinforcement learning with verifiable rewards for cross-cultural entity translation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5616–5638, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.254. URL https://aclanthology.org/2026.acl-long.254/.

Leqi Zhu, Junyan Ye, Kaiqing Lin, Zhiyuan Yan, Conghui He, and Weijia Li. Fakevlm-r1: Internalizing physical laws via cot for synthetic image detection, 2026a. URL https://arxiv.org/abs/2605.30062.

Xuekang Zhu, Xiaochen Ma, Lei Su, Zhuohang Jiang, Bo Du, Xiwen Wang, Zeyu Lei, Wentao Feng, Chi-Man Pun, and Ji-Zhe Zhou. Mesoscopic insights: orchestrating multi-scale & hybrid architecture for image manipulation localization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 11022–11030, 2025.

Xuekang Zhu, Ji-Zhe Zhou, Kaiwen Feng, Chenfan Qu, Xiwen Wang, Yunfei Wang, Liting Zhou, and Jian Liu. Revisiting image manipulation localization under realistic manipulation scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7198–7207, 2026b.

A License

All datasets and models used in our experiments are available for research purposes. We declare that this study is conducted solely for academic research and has no commercial purpose.

B Experimental Setup

The training of TwSG is conducted in two distinct phases: (1) Cold-start SFT. To initialize the reasoning capability of the model, we perform SFT with a learning rate of 1e-5 and a global batch size of 32; (2) Following the cold-start SFT, we employ TL-GRPO to further refine the model's reasoning trajectory. Unlike SFT, this stage emphasizes final outcome correctness rather than dense CoT supervision. We set the reward coefficient β to 0.0 and utilize a dual-clip epsilon strategy with $\epsilon = 3 \times 10^{-4}$ and $\epsilon_{high} = 4 \times 10^{-4}$ to stabilize importance sampling. The learning rate is decayed to 1×10^{-6} with a warmup ratio of 0.05. For each prompt, we sample $G = 8$ generations to compute the relative reward. The models are trained on a cluster of 128 NVIDIA A100 GPUs (80GB). The total computational budget for training and evaluation ranges from 50 to 120 wall-clock hours, depending on the backbone scale.

C Compare with close-source MLLMs

As shown in Table 3, on CharXiv-R (Wang et al., 2024b), TwSG-8B achieves **67.8%**, outperforming strong proprietary models such as GPT-4.1 (56.7%) and GPT-4.5 (55.4%), and even surpassing large-scale models like Qwen3-VL-235B-A22B-Thinking (66.1%) despite a significantly smaller parameter size. Compared with recent "thinking with images" models, our method demonstrates superior performance without explicit test-time reasoning or tool use, indicating that our gains come from improved intrinsic multimodal reasoning rather than inference-time scaling.

Table 3: Performance comparison on chart reasoning benchmarks. Left: results on ChartQA with open-source MLLMs. Right: comparison with proprietary models on the CharXiv-R subset. "Tools" indicates whether external tool use is enabled. **Bold** denotes the best, and underline the second best. ChartMOE performs inference with an external Python tool.

(a) Comparison with general MLLMs on ChartQA.

Model	Venue	Tools	ChartQA (%)
General-domain MLLMs			
Qwen2.5-VL-7B (Bai et al., 2025b)	-		87.3
Qwen3-VL-8B (Bai et al., 2025a)	-		85.6
Thyme (Zhang et al., 2025)	ICLR'26	✓	86.1
DeepEyesV2 (Hong et al., 2026)	ICLR'26	✓	<u>88.4</u>
VGR-7B (Wang et al., 2026)	ICLR'26	✓	72.8
CvTR-domain MLLMs			
ChartAst-S (Meng et al., 2024)	ACL'24		79.9
ChartLLaMa (Han et al., 2023)	-		69.7
BigCharts-R1-7B (Masry et al., 2025b)	COLM'25		89.8
Chart-R1 (Chen et al., 2025)	-		91.0
ChartMOE* (Xu et al., 2025a)	ICLR'25	✓	87.8
Ours (TwSG-8B)	-		**92.9**

(b) Comparison with large size MLLMs on CharXiv-R.

Model	Provider / Venue	CharXiv-R (%)
General-domain MLLMs		
GPT-4o	OpenAI	47.1
Qwen3-VL-8B-Thinking	Alibaba	53.0
GPT-4.5	OpenAI	55.4
GPT-4.1	OpenAI	56.7
Qwen3-VL-235B-A22B-Instruct	Alibaba	62.1
Qwen3-VL-32B-Instruct	Alibaba	62.8
Qwen3-VL-32B-Thinking	Alibaba	<u>65.2</u>
Qwen3-VL-235B-A22B-Thinking	Alibaba	66.1
CvTR-domain MLLMs		
BigCharts-R1-7B (Masry et al., 2025b)	COLM'25	41.3
Bespoke-MiniChart (Meng et al., 2025)	-	45.4
Chart-R1 (Chen et al., 2025)	-	46.2
Ours (TwSG-8B)	-	**67.8**

D Verify Reward

Evaluating the quality of open-ended intermediate reasoning steps is notoriously difficult, as traditional string-matching metrics often fail to assess semantic correctness and visual faithfulness. To address this limitation, we leverage an expert MLLM as an automated judge to compute the verify reward (r_{verify}). This follows the general practice of LLM/MLLM-as-a-judge evaluation, where strong foundation models are used to assess open-ended outputs beyond exact string matching.

Table 4: Ablation studies on reward verification and reward weighting.

(a) Judge Model

Method	ChartQAPro	TableVQA-Bench
GPT-4o	54.44	85.32
Qwen3.6-plus	56.32	85.90
Qwen3-VL-72B-Instruct (Default)	55.60	86.79

(b) Reward Weights

λ_1	λ_2	λ_3	ChartQAPro	TableVQA-Bench
0.00	1.00	0.00	52.30	82.91
0.05	0.90	0.05	54.83	85.74
0.10	0.80	0.10	55.60	86.79

Table 5: Impact of decoding hyperparameters. Left: different random seeds. Right: different sampling temperatures.

(a) Seed

Seed	ChartQAPro	TableVQA
42	55.60	86.79
1024	55.47	86.62
8192	55.71	86.84

(b) Temperature

Temp	ChartQAPro	TableVQA
0.1	55.60	86.79
0.5	55.42	86.51
0.9	55.18	86.35

The primary advantage of this approach is its ability to perform nuanced, visually grounded evaluation. By framing the assessment as a rigorous multiple-choice task, the expert MLLM can explicitly differentiate critical visual hallucinations from acceptable minor phrasing variations.

D.1 Ablation Experiment

Effect of judge models. Table 4 (a) shows that the choice of judge MLLM for r_{verify} has only a marginal impact on the final performance. Replacing the default Qwen3-VL-72B-Instruct (Bai et al., 2025a) with stronger proprietary judges does not lead to consistent gains: GPT-4o obtains 54.44 on ChartQAPro and 85.32 on TableVQA-Bench, while Qwen3.6-plus (Qwen Team, 2026) reaches 56.32 and 85.90, respectively. In comparison, our default Qwen3-VL-72B-Instruct achieves 55.60 on ChartQAPro and the best result of 86.79 on TableVQA-Bench. The overall variation across different judges is small, suggesting that our reward verification is not highly sensitive to the specific judge model once the judge has sufficient multimodal reasoning capability. From a cost perspective, using Qwen3-VL-72B-Instruct is also more practical.

Effect of reward weights. Table 4 (b) analyzes the effect of different reward-weight configurations in TL-GRPO. When only the answer reward is used, i.e., $(\lambda_1, \lambda_2, \lambda_3) = (0.00, 1.00, 0.00)$, the model obtains 52.30 on ChartQAPro and 82.91 on TableVQA-Bench, indicating that direct answer supervision provides the main optimization signal. After introducing small weights for the format reward and verification reward, the performance improves to 54.83 and 85.74 with $(0.05, 0.90, 0.05)$, showing that output-format regularization and judge-based verification provide complementary guidance. The default setting $(0.10, 0.80, 0.10)$ achieves the best performance on both benchmarks, with 55.60 on ChartQAPro and 86.79 on TableVQA-Bench. This suggests that answer correctness should remain the dominant reward, while lightweight format and verification rewards are beneficial for stabilizing the reasoning process and improving final accuracy.

Effect of decoding hyperparameters. Table 5 evaluates the robustness of our method under different decoding hyperparameters. Across three random seeds, the performance remains stable, with only minor fluctuations on both ChartQAPro and TableVQA-Bench. This indicates that the reported results are not sensitive to random sampling effects. We also vary the sampling temperature from 0.1 to 0.9. The performance changes only slightly, while higher temperatures lead to a small degradation due to increased output randomness. These results suggest that our method is robust to decoding configurations and does not rely on carefully tuned test-time sampling.

```
System Instruction:
You are an expert visual evaluator. Your task is to carefully assess the accuracy of a generated
"Observation" based strictly on the provided image. Compare the textual description with the visual
evidence and determine its level of accuracy.
Input:
- Image: [Sub Image]
- Observation: [Visual Description]
Evaluation Options:
Please select the most appropriate option from the following categories:
    A. Completely Accurate: The observation perfectly aligns with the visual content, containing
       absolutely no hallucinations or factual errors.
    B. Mostly Accurate: The observation is largely correct and visually grounded, with only minor or
       negligible discrepancies that do not impact the overall understanding.
    C. Partially Accurate: The observation contains a mix of accurate visual descriptions and
       noticeable hallucinations or errors.
    D. Mostly Inaccurate: The observation is dominated by severe hallucinations, or fundamentally
       contradicts the primary visual evidence.
    E. Completely Unrelated: The observation completely fails to describe the image, or is entirely
       irrelevant to the visual content.
Output Format:
Provide your judgment by outputting ONLY a single uppercase letter corresponding to your choice (A, B,
C, D, or E). Do not include any explanations, punctuation, or extra text.
```

Figure 4: Prompt for Verify Reward.

D.2 API Cost

For the exact prompt template used to query the expert MLLM, including the detailed instructions and the complete option descriptions provided to the model, please refer to Figure 4.

Public API pricing shows that GPT-4o is substantially more expensive (OpenRouter, 2026), commonly listed around \$2.50/M input tokens and \$10.00/M output tokens, while Qwen3.6-plus is listed around \$0.355/M input tokens and \$1.99/M output tokens. By contrast, Qwen-VL/Qwen3-VL family models are available at lower-cost API endpoints, and Qwen3-VL-72B-Instruct can also be deployed locally as an open-weight judge, further reducing large-scale training-time annotation cost. Therefore, we adopt Qwen3-VL-72B-Instruct as the default judge since it provides a better trade-off between verification quality and training cost.

E Cold-Start SFT Data Construction

This section details the construction process of the Cold-Start SFT data, including the distillation of CvTR-domain datasets and their statistical distributions. The primary objective of our distillation process is to extract interleaved CoT from diverse data sources, enabling the MLLM to internalize this reasoning paradigm—a strategy that has been widely adopted in recent state-of-the-art methodologies (Xia et al., 2025; Guo et al., 2025). Consequently, this phase focuses on constructing a high-quality offline dataset where the inputs comprise the image I, the question Q, and the ground truth label from the original training set, while the output is the corresponding elicited CoT.

We collect raw training data from the original training splits of TableQA, TableQA-X, and Visual-TableQA. To further enrich the visual layouts, especially for multi-chart and multi-table scenarios, we randomly synthesize composite samples consisting of two charts, two tables, or one chart and one table, while preserving the original question-answer pairs. This strategy allows us to efficiently expand the coverage of complex chart-table layouts without introducing additional annotation cost. We then perform the following filtering procedure:

Step 1: ROI Identification. The initial step involves leveraging a leading closed-source MLLM to identify all potential ROI bounding boxes. Specifically, given the full image I and the question Q, the model determines

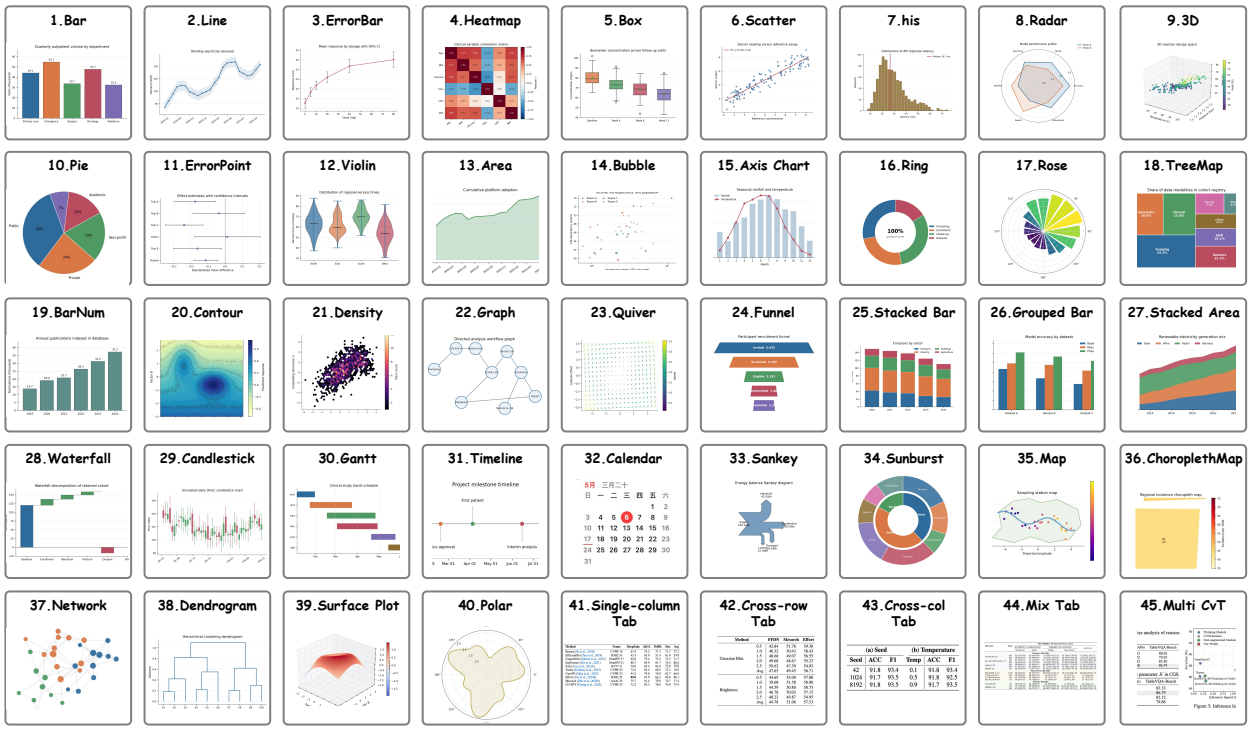

Figure 5: Illustrative examples of the 45 chart and table categories covered in our benchmark.

which sub-regions are most relevant to the query and returns their absolute coordinates. For this ROI generation, we utilize GPT-4o. The detailed prompt for this task is illustrated in Figure 6.

Step 2: Region Description. In the second step, we crop the corresponding sub-regions based on each generated bounding box to obtain sub-images sub_I. We then prompt the MLLM to generate a comprehensive description for each region. Unlike standard global description prompts, our instruction is explicitly optimized for text recognition within charts and tables (as detailed in Figure 7). One area exects once time. We similarly employ GPT-5.1 for this high-fidelity content description task.

Step 3: CoT Refinement and Structuring. The third step refines the CoT generated in the preceding stages. We empirically observed that raw CoT often contains colloquialisms and excessive redundancy. Given the critical role of data quality, we instruct the MLLM to perform two key tasks: (1) simplify the reasoning path by preserving only essential text-related descriptions, and (2) extract textual content strictly pertinent to the question. For instance, if a query specifically targets a "yellow region," any extraneous global descriptions are distilled to retain only information concerning that target area. We require the model (Qwen3.6-Plus (Qwen Team, 2026)) to produce this refined data in a structured JSON format. Finally, we encapsulate the comprehensive visual observations within `<observation>` tags, while the question-specific reasoning is enclosed within `<think>` tags, as shown in Figure 8.

Finally, we conduct a hallucination detection phase for each data entry to ensure faithfulness. The evaluation methodology and the corresponding verification prompt are provided in Figure 4. Unlike the main text, this section details the specific, step-by-step construction process.

Figure 5 illustrates the diversity of visualization and table types covered in our benchmark. The benchmark includes 45 categories, spanning standard statistical charts, distribution plots, relational diagrams, temporal visualizations, geographic maps, hierarchical structures, and complex tables. Specifically, it covers basic chart types such as bar, line, scatter, pie, histogram (his), heatmap, box, violin, radar, area, bubble, and 3D plots; advanced visualization forms such as error bars, error points, multi-axis charts (Axis Chart), ring charts (Ring), rose charts (Rose), treemaps, contour plots, density plots, quiver plots, funnels, stacked/grouped bars, stacked areas, waterfalls, candlesticks, Gantt charts, timelines, calendar heatmaps, Sankey diagrams,

```
You are a visual localization assistant. Given a question and an input image, identify the region(s)
of interest (ROIs) that contain the visual evidence needed to answer the question.

Do not answer the question. Return only the relevant bounding box(es) and a brief description of what
each region contains.

Return the result as a JSON list of dictionaries only:

[
  {
    "bbox_2d": [x1, y1, x2, y2],
    "region_summary": "A brief description of the localized region."
  },
  ...
]

Requirements:

     1. Each dictionary corresponds to one question-relevant ROI.

     2. The bounding box must tightly cover the relevant region.

     3. The coordinates must follow the format [x1, y1, x2, y2].

     4. The region summary should be brief and factual.

     5. Do not extract detailed text or numbers at this stage.

     6. Do not provide the final answer.

     7. Do not hallucinate regions that are not visible in the image.

Question: {question}
Image: {image}
```

Figure 6: Prompt for ROI extraction.

```
You are a visual reasoning assistant.
Given a image summary and an input image, identify the image region that contains information useful
for answering the question.
Describe the given region and explicitly include the key visible text/numbers from that region.
Return JSON only:

{
  "bbox_2d": [x1, y1, x2, y2],
  "area_description": "Describe the given region, including its spatial context and all
  question-relevant visible text/numbers exactly as shown in the image."
}

Requirements:

     1. Include exact visible text/numbers in area_description.

     2. Include labels, row/column names, cell values, axis values, or legends if relevant.

     3. List multiple relevant values explicitly.

     4. Do not hallucinate missing text.

Input:
Image Summary: {region_summary}
Image: {image}
```

Figure 7: Prompt for region description and text extraction.

sunburst charts, maps, choropleth maps, network graphs, dendrograms, surface plots, and polar plots. In addition, the benchmark includes five table-oriented categories: single-column tables, cross-row tables, cross-column tables, mixed tables, and Multi chart and table layouts (Multi CvT). Following the query design of ChartQAPro (Masry et al., 2025a), we define eight question categories in our benchmark, including Mathematical Reasoning, Visual Reasoning, Conversational, Multiple-Choice, Hypothetical, Fact-Checking, Unanswerable, and Multi-Chart QA. To build a balanced benchmark, we first train a local Qwen3-32B

```
You are a CoT refinement assistant. Given a question, an input image, and raw textual descriptions
extracted from multiple localized regions, clean and structure the reasoning information for
high-quality data construction.
Your task is to simplify the raw descriptions and retain only the information that is directly useful
for answering the question. Remove colloquial expressions, redundant descriptions, irrelevant global
context, and any content unrelated to the question. For each localized region, output a concise
question-relevant textual description.
Return JSON only:

{
    "cleaned_description": "A concise description of the question-relevant
    text, numbers, labels, or visual evidence in this region."
}

Requirements:

    1. Preserve only essential information that is directly related to the question.

    2. Extract textual content strictly pertinent to the question.

    3. Remove redundant, colloquial, or irrelevant descriptions.

    4. If the question targets a specific region, such as a yellow region, keep only information
       about that target region.

    5. Do not add information that is not visible in the image or not present in the raw text.

    6. Do not answer the question.

    7. The output must be a JSON dictionary.

Input:
Question: {question}
Sub Image: {image}
Raw region descriptions: {raw_region_descriptions}
```

Figure 8: Prompt for CoT refinement and structuring.

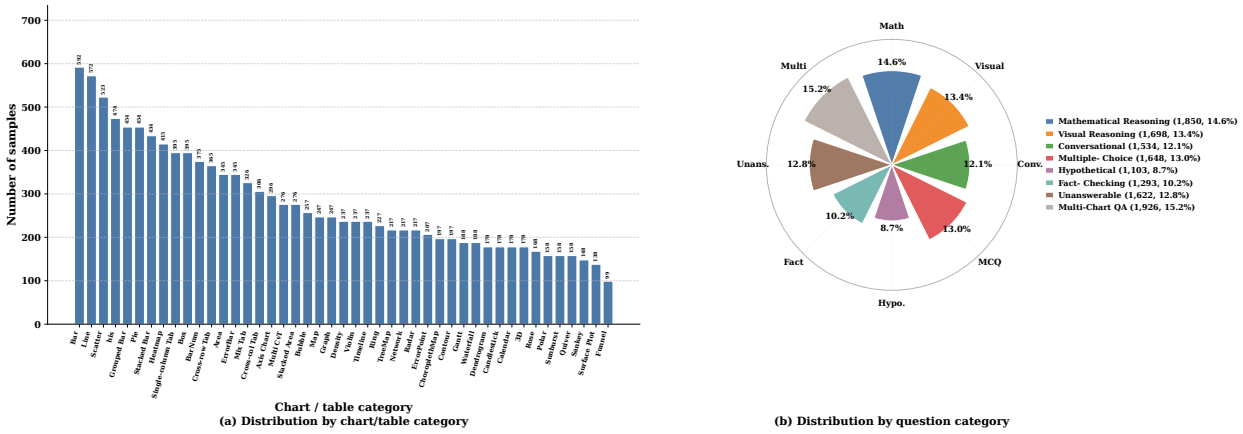

Figure 9: The Static of TwSG-12K.

classifier to categorize all candidate questions in the seed pool, and further train a Qwen3-VL-8B classifier to identify the chart/table type of each sample. Based on these automatic annotations, we perform category balancing and then manually filter redundant or low-quality samples. As shown in Figure 9, the final benchmark contains 12,674 samples spanning 45 chart/table categories and 8 question categories. The chart and table distribution exhibits a mildly long-tailed pattern, where common categories such as bar, line, scatter, and pie occupy relatively larger portions, while rarer categories remain sufficiently represented. Meanwhile, the question distribution is relatively balanced across the eight categories, which helps ensure comprehensive evaluation over diverse reasoning skills.

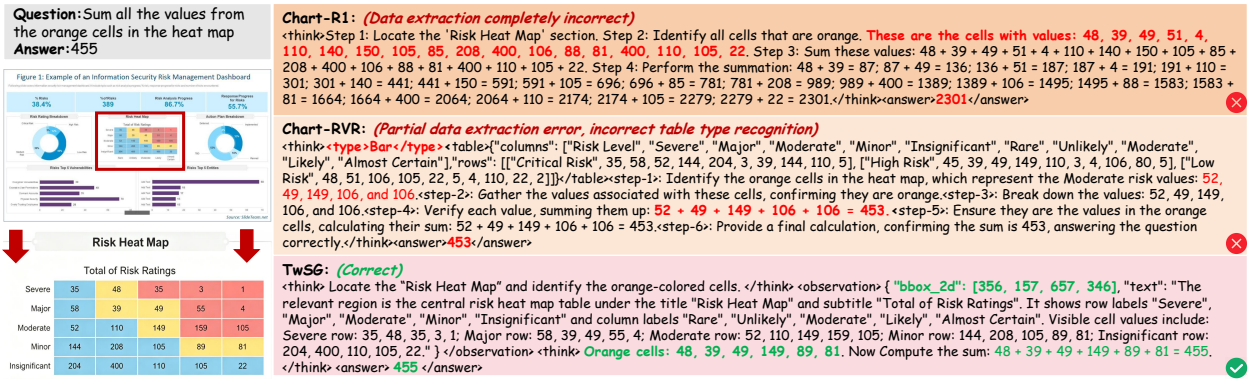

Figure 10: Reasoning comparisons between TwSG and existing CvTR-domain MLLMs.

The Distillation Objective. Given the teacher-generated interleaved trajectory τ, we optimize the student model M_θ using a standard cross-entropy loss:

$$\mathcal{L}_{TwSG} = -\sum_{j=1}^{|\tau|} \log P(x_j | I, Q, x_{<j}; \theta) \tag{5}$$

where x_j are the tokens in the interleaved sequence. By training on these trajectories, the student learns to internally simulate the "Reason-Observation-Reason" process.

This approach mimics the reasoning process of iteratively referencing visual charts during complex, multi-step problem-solving. Consequently, it significantly mitigates the "coordinate hallucination" prevalent in MLLMs (Fu et al., 2025)—instances where the predicted coordinates decouple from the reasoning context. Furthermore, by integrating an expert teacher model, our method effectively alleviates the inherent hallucinations typically observed during the model's CoT.

F Qualitative Analysis

As shown in Figure 10, existing CvTR-domain MLLMs can identify the relevant heat-map region but often fail to establish a faithful mapping between color-coded cells and their textual values. Chart-R1 directly extracts incorrect values from non-orange cells, while Chart-RVR repeats the same error despite producing a longer reasoning trace, suggesting that extended reasoning alone does not guarantee reliable visual grounding. In contrast, TwSG first localizes the key region through the ⟨observation⟩ field and explicitly lists the orange-cell values before aggregation.