

Chart-RL: Generalized Chart Comprehension via Reinforcement Learning with Verifiable Rewards

Xin Zhang, Xingyu Li, Rongguang Wang, Ruizhong Miao,
Zheng Wang, Yuying Wang, Dan Roth, Chenyang Li

Oracle AI

{xin.j.zhang, xingyu.li, rongguang.wang, ruizhong.miao, z.zheng.wang, yuying.wang, dan.roth, chenyang.li}@oracle.com

Abstract

Accurate chart comprehension represents a critical challenge in advancing multimodal learning systems, as extensive information is compressed into structured visual representations. However, existing vision-language models (VLMs) frequently struggle to generalize on unseen charts because it requires abstract, symbolic, and quantitative reasoning over structured visual representations. In this work, we introduce Chart-RL, an effective reinforcement learning (RL) method that employs mathematically verifiable rewards to enhance chart question answering in VLMs. Our experiments demonstrate that Chart-RL consistently outperforms supervised fine-tuning (SFT) across different chart understanding benchmarks, achieving relative improvements of 16.7% on MultiChartQA, and 11.5% on ChartInsights. We conduct robustness analysis, where Chart-RL achieves enhanced performance in 18 of 25 perturbed chart categories, demonstrating strong consistency and reasoning capability across visual variations. Furthermore, we demonstrate that task difficulty and inherent complexity are more critical than data quantity in RL training. For instance, Chart-RL trained on merely 10 complex chart-query examples significantly outperforms models trained on over 6,000 simple examples. Additionally, training on challenging reasoning tasks not only improves in-domain generalization relative to simpler tasks, but also facilitate strong transfer to out-of-domain visual mathematical problems.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) has demonstrated remarkable success in enhancing the reasoning capabilities of LLMs (Guo et al., 2025; Jaech et al., 2024; Team et al., 2023). However, one critical area that remains relatively underexplored is chart understanding in VLMs, whose challenges such as symbolic reasoning and precise data extraction distinguish it fundamentally

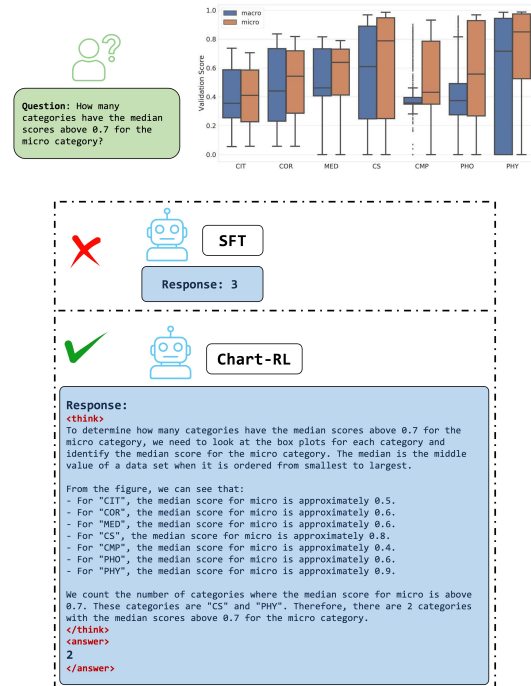


Figure 1: Comparison of Chart-RL vs SFT to reason on complex chart question answering.

from traditional vision-language tasks. Unlike natural images, charts exhibit extreme diversity in types (bar charts, pie charts, scatter plots, etc.) and structural organizations, with performance heavily dependent on the training data distribution (Huang et al., 2024). Current VLMs often struggle to bridge the gap between extracting descriptive information and deriving correct answers that require multi-step reasoning (Kamath et al., 2023).

Recent studies to improve chart question answering have explored various strategies, such as converting charts into tabular formats suitable for prompting LLMs (Liu et al., 2022), instruction tuning on multiple chart-related objectives (Schleid et al., 2025), and generating synthetic chart data to augment VLM’s capability in real-world chart comprehension (Huang et al., 2025a). Nonetheless,

the majority of chart comprehension methods rely on either larger curated datasets that cover broader chart and question types, or SFT using domain-specific datasets. While large-scale datasets such as ChartQA (Masry et al., 2022), PlotQA (Methani et al., 2020), and FigureQA (Kahou et al., 2017) have significantly expanded coverage of chart types and reasoning scenarios, they often focus on specific chart formats with template-based questions that may not fully capture real-world complexity. Although domain-specific SFT can enhance model performance on targeted task categories, it frequently leads to catastrophic forgetting and performance degradation on untrained tasks due to data distribution shifts (Luo et al., 2023; Kalajdziewski, 2024). This phenomenon fundamentally limits the capability of current approaches to generalize across diverse chart types and reasoning scenarios in real-world applications (Schulze Buschoff et al., 2025; Huang et al., 2023).

To address these limitations, we propose **Chart-RL**, an efficient reinforcement fine-tuning framework that employs R1-style reinforcement learning to enhance chart understanding capabilities in VLMs. A key insight motivating our approach is that chart-based questions frequently correspond to mathematically verifiable ground truths, making them well-suited for rule-based reward functions in RL training. More importantly, many charts encapsulate sufficiently complex visual and semantic information that necessitate iterative reasoning processes. We find that such reasoning complexity is of central importance because it enables models to develop advanced reasoning capabilities that can’t be acquired through simple chart interpretation tasks (Parashar et al., 2025). Leveraging group relative policy optimization (GRPO) with chart-aligned accuracy and format rewards (Shao et al., 2024; Shen et al., 2025), we train VLMs to develop robust reasoning capabilities that generalize across different chart types and complexity levels. Our approach circumvents the limitation of conventional SFT methods, where task-specific adaptation results in narrow specialization rather than transferable reasoning skills (Luo et al., 2023).

As illustrated in Figure 1, we demonstrate that Chart-RL improves performance across different chart understanding tasks, ranging from simple data extraction to complex multi-step reasoning, while achieving remarkable data efficiency. By training on complex chart problems, Chart-RL develops proficiency in low-level component tasks

without requiring explicit instruction tuning.

The key contributions of this paper include:

- **RLVR for Task-Agnostic Chart Comprehension:** We introduce Chart-RL, a reinforcement fine-tuning strategy that uses reinforcement learning with verifiable rewards to enhance generalized chart comprehension capabilities in VLMs, achieving significant improvements over SFT approaches.
- **Data-Efficient RL Training:** We demonstrate, for the first time, that RL training on complex chart reasoning tasks enables robust generalization, achieving consistent improvements across diverse benchmarks without task-specific optimization.
- **Comprehensive Experimental Validation of Robustness and Generalization:** Our extensive empirical evaluation reveals a critical insight: **task complexity**, rather than the quantity of training data, is pivotal in developing generalizable chart understanding capabilities.

2 Related Works

2.1 Chart-Based Question Answering

Chart understanding has evolved from multi-stage pipelines (Masry et al., 2022) to end-to-end models. This transformation was pioneered by OCR-free architectures like Donut (Kim et al., 2022), which paved the way for powerful VLMs. Recent progress has been predominantly achieved through SFT approaches, with models such as Pix2Struct (Lee et al., 2023) and UniChart (Masry et al., 2023) being adapted for direct chart-to-answer generation. Concurrently, advanced models including LLaVA (Liu et al., 2023), PaliGemma (Beyer et al., 2024), and ChartLlama (Han et al., 2023) have been fine-tuned to address more sophisticated and diverse reasoning challenges.

Despite its success in task-specific adaptation, SFT faces critical limitations in the domain of chart understanding. The need for large, meticulously annotated datasets makes the process costly and prone to overfitting, where models learn to exploit data artifacts rather than acquiring robust reasoning principles. Furthermore, SFT-based models often exhibit poor generalization ability: performance gains on one type of task do not reliably transfer to related tasks or different data distributions (Schulze Buschoff et al., 2025).

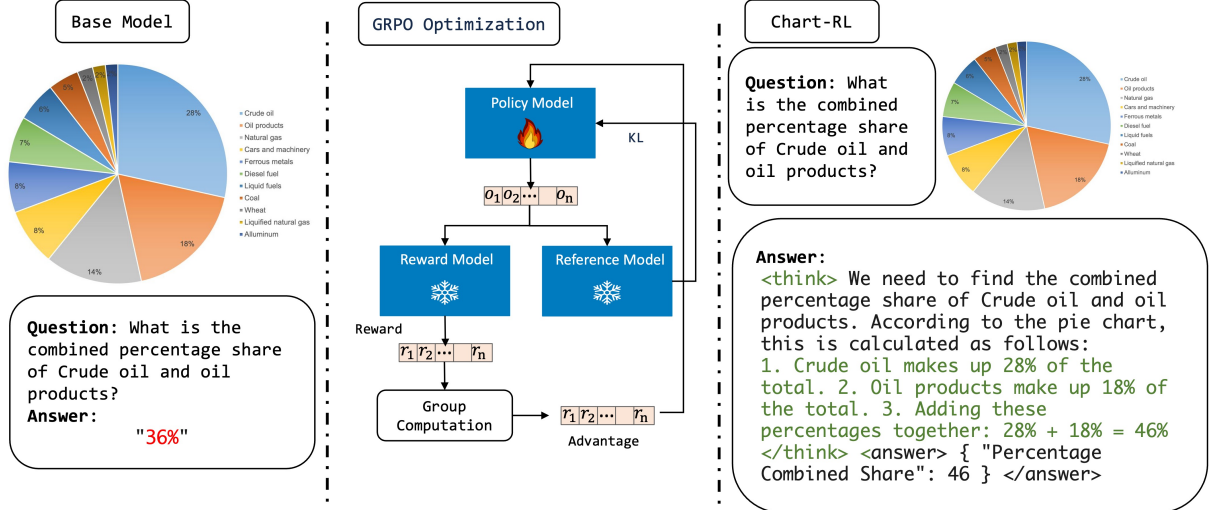


Figure 2: Schematic diagram of Chart-RL framework with GRPO optimization for chart comprehension.

2.2 Reinforcement Fine-Tuning in VLMs

The success of R1-style reinforcement learning in enhancing LLM reasoning capabilities (Jaech et al., 2024; Guo et al., 2025) has inspired extensive research into its application to VLMs. For example, R1-OneVision (Yang et al., 2025) and Vision-R1 (Huang et al., 2025b) construct multimodal reasoning datasets by converting visual information into textual format before RL training, while other studies focus on specific applications such as object counting (Chen et al., 2025) or employ staged training strategies with progressive difficulty rewards (Peng et al., 2025; Deng et al., 2025). Notably, several studies have observed the “visual aha moment” in VLMs (Zhou et al., 2025; Meng et al., 2025), demonstrating emergent reasoning capabilities. The VLM-R1 framework (Shen et al., 2025) provides a comprehensive investigation into applying RL to general visual understanding tasks, including Referring Expression Comprehension (REC) and Open-Vocabulary Object Detection (OVD).

Despite these advances, the reasoning capabilities cultivated by existing R1-style VLMs are primarily tailored for understanding of natural images. Since charts present highly condensed information in diverse structural formats, the unique challenges of chart understanding have yet to be systematically addressed. These tasks require interpreting relationships, performing calculations, and synthesizing information from graphical elements like bars, lines, and legends. Building on the principles established in the VLM-R1 (Shen et al., 2025), we propose an R1-type reinforcement fine-tuning ap-

proach specifically for chart understanding, aiming to instill the robust generalization and reasoning capabilities.

3 Methods

3.1 Chart-RL Framework

We present Chart-RL, an extension of the open-source VLM-R1 platform (Shen et al., 2025) that adapts R1-style reinforcement fine-tuning for chart understanding tasks. As shown in Figure 2, we implement a verifiable reward function that produces a score based on alignment between model outputs and annotated ground truths. In addition, a binary format reward function is employed to enforce task-specific output structures. This modular reward design enables Chart-RL to adapt to different chart-related training tasks with mathematically verifiable annotations.

3.2 Accuracy and Format Rewards

In Chart-RL, we employ the GRPO algorithm, which evaluates a group of candidate responses to guide policy updates, eliminating the need for a separate critic (Shao et al., 2024). The GRPO algorithm begins by sampling N candidate responses $\{o_1, o_2, \dots, o_N\}$ from the current policy π_θ for a given query q . Each response o_i is scored using a reward function $R(q, o_i)$, and the resulting rewards r_i are normalized to compute the advantage A_i for each response as:

$$A_i = \frac{r_i - \text{mean}\{r_1, r_2, \dots, r_N\}}{\text{std}\{r_1, r_2, \dots, r_N\}}, \quad (1)$$

where A_i measures the relative advantage of response i compared to other responses in the sampled group. The policy π_θ is then updated to increase the likelihood of generating responses with higher advantages using the following objective:

$$\begin{aligned} \mathcal{J}_{GRPO}(\theta) &= \mathbb{E}_{\{o_i\} \sim \pi_{\theta_{old}}(q)} \\ &\frac{1}{N} \sum_{i=1}^N \{ \min[s_1 \cdot A_i, s_2 \cdot A_i] - \beta \mathbb{D}_{KL}[\pi_\theta || \pi_{ref}] \} \\ s_1 &= \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)} \\ s_2 &= \text{clip} \left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1 - \epsilon, 1 + \epsilon \right) \end{aligned} \quad (2)$$

Accuracy Reward. To ensure verifiable training signals, we sample query-chart pairs that produce mathematically deterministic results for RL training. Our accuracy reward design is based on the relative error between the model’s predicted result v_p and the ground-truth solution v_g , evaluated against a predefined precision threshold τ . The accuracy reward score $S(v_p, v_g)$ is computed as:

$$S(v_p, v_g) = \begin{cases} 1, & \text{correct output} \\ 0, & \text{incorrect or malformed output} \end{cases} \quad (3)$$

Format Reward. We employ a binary format reward to enforce adherence to the prescribed output structure. This reward function assigns a value of 1.0 if the model’s response strictly follows the required template format, and 0 otherwise. Specifically, the model must structure its output with two distinct components: (1) a reasoning process enclosed within `<thinking>...</thinking>` tags, and (2) a final answer in JSON format enclosed within `<answer>““ json...““</answer>` tags.

To implement this structured output format, we utilize carefully designed prompts during Chart-RL training as shown below. The system prompt establishes the conversational framework and defines the expected output structure, while the generation prompt reinforces these formatting requirements for each specific query.

System 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 `<thinking> </thinking>` and `<answer> </answer>` tags, respectively, i.e., `<thinking> reasoning process here </thinking><answer> answer here </answer>`

4 Experiments

4.1 Training & Evaluation Setup

Baseline VLMs. We use Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct as our baseline models. Recent studies show that small language models (SLMs) are sufficiently powerful and, when adapted with efficient post-training, provide a cost-effective and therefore more promising foundation for practical agentic systems (Belcak et al., 2025). In addition, these models align better with our practical development needs. For our ablation study, we focus specifically on the Qwen2.5-VL-3B-Instruct model to systematically analyze the effects of different fine-tuning strategies.

Training Datasets. We conduct RL training on two types of chart-based tasks:

- **Easy Task:**

- This task requires only single-step reasoning and involves straightforward chart question-answering, such as direct numeric extraction from a single chart element.
- We use 6,200 subsampled charts from PlotQA (Methani et al., 2020), focusing on straightforward numerical queries.

- **Hard Task:**

- This task involves multi-step reasoning by performing a sequence of simple sub-tasks across one or multiple charts, ultimately producing responses through the aggregation of intermediate results.
- From the CharXiv validation set (Wang et al., 2024), we subsample 448 charts with mathematically verifiable ground truth, tailored for multi-hop reasoning in Chart-RL training.

Unlike the easy tasks that focus on straightforward numerical extraction, the hard tasks require multi-hop reasoning across chart elements, making

Model Name	Method	MultiChartQA		ChartInsights	
		Overall		Reasoning	
		Accuracy	Rel. Δ	Accuracy	Rel. Δ
Qwen-2.5VL-3B	Baseline	35.2	–	36.4	–
	SFT	32.9	-6.6%	34.7	-4.7%
	CoT-SFT	<u>39.1</u>	+11.1%	<u>40.3</u>	+10.7%
	Chart-RL(ours)	41.1	+16.7%	40.6	+11.5%
Qwen-2.5VL-7B	Baseline	47.5	–	48.4	–
	SFT	46.5	-2.1%	46.3	-4.3%
	CoT-SFT	<u>51.1</u>	+7.6%	<u>49.9</u>	+3.1%
	Chart-RL(ours)	55.9	+17.7%	50.6	+4.5%

Table 1: Evaluation results on two benchmarks comparing our Chart-RL approach against baseline, SFT, and CoT-SFT on hard task.

the 448-chart training subset sufficiently complex for RL training. Unless otherwise noted, all reported Chart-RL results use query-chart pairs from this 448-chart hard task training set. We provide examples of easy and hard tasks in Appendix A.2.

Training Hyper-parameters. We employ the VLM-R1 (Shen et al., 2025) pipeline using our customized accuracy and formatting rewards. All RL experiments are conducted on 8 NVIDIA H100 GPUs. At each update, we generate 8 samples per input prompt. We set per-device batch size to 1 and accumulate gradients over 4 steps. We apply parameter-efficient fine-tuning (PEFT) via LoRA (Hu et al., 2022), setting the adapter with rank $r = 64$, scaling factor $alpha = 128$, and dropout rate of 0.05. To enable joint vision-language adaptation, we keep all vision modules unfrozen throughout RL fine-tuning.

Evaluation Benchmarks. We evaluate Chart-RL on three state-of-the-art chart comprehension benchmarks, including MultiChartQA (Zhu et al., 2024), ChartInsights (Wu et al., 2024), and RobustCQA (Mukhopadhyay et al., 2024).

- **RobustCQA:** Measures the robustness of VLMs under systematically perturbed chart variations.
- **MultiChartQA:** Assesses the model’s ability to perform multi-hop reasoning and to integrate information across multiple charts.
- **ChartInsights:** Evaluates fine-grained analytical performance across seven chart types and ten task categories.

Category	Dataset	# Charts	# Queries
Training	CharXiv	448	448
	PlotQA	6,200	6,200
Evaluation	RobustCQA	400	400
	MultiChartQA	1,370	2,000
	ChartInsights	2,000	22,000

Table 2: Statistics of training and evaluation datasets.

CoT Trajectory Generation. We use GPT-4o as a teacher model to process query-chart pairs from the hard task and generate validated reasoning traces. The resulting CoT trajectories then serve as supervised fine-tuning data. To compare CoT-SFT against RL training, the complete prompt template is provided in the Appendix A.1.

4.2 Fine-Tuning Strategy Comparison

We evaluate our proposed Chart-RL against two baseline models and two fine-tuning paradigms: standard Supervised Fine-Tuning (SFT) and Chain-of-Thought SFT (CoT-SFT). As shown in Table 1, Chart-RL outperforms all baselines and SFT variants across both chart understanding benchmarks. Specifically, standard SFT yields a performance degradation on both datasets, suggesting that conventional, task-agnostic tuning is insufficient for capturing the complex reasoning required for chart question answering. Incorporating intermediate reasoning steps through CoT-SFT leads to improvements over both the baseline and standard SFT. However, Chart-RL yields the most substantial gains. Despite not being explicitly tailored to

these evaluation benchmarks, Chart-RL exhibits strong zero-shot generalization on the 3B model, achieving relative improvements of 16.7% on MultiChartQA and 11.5% on ChartInsights over the baseline.

We attribute Chart-RL’s generalization and reasoning capabilities to the exploration inherent in RL training (Peng et al., 2025). These results demonstrate that our Chart-RL framework provides a robust approach for advancing chart understanding tasks.

4.3 Robustness Analysis

We evaluate Chart-RL’s robustness across 25 perturbed chart categories using complex charts and questions from the RobustCQA dataset, which systematically assesses VLMs’ consistency across different visual representations of identical underlying data. As shown in Table 3, Chart-RL demonstrates improved performance in 18 of 25 categories (72%), compared to the SFT. In contrast, SFT only achieves better results in only 2 categories (8%). The remaining 5 categories exhibit identical performance (20%).

The most significant improvements occur in categories involving chart layout modifications and visual styling, such as hatching patterns, legend positioning, and tick orientation. These improvements suggest that reinforcement learning with chart-specific reward signals likely enhances generalization capabilities, making the model robust to layout and structural variations.

4.4 Data Efficiency in RL Training

Inspired by recent work on RL training efficiency (Wang et al., 2025), we examine the training dynamics using minimal training examples. Specifically, we aim to investigate how few verifiable-reward samples are required to achieve efficient visual reasoning in chart. Figure 3 compares Chart-RL’s training dynamics with limited training data, where the number of charts is restricted to 10, 100, or 448 randomly subsampled instances of equivalent task complexity. Notably, with only 10 samples, the accuracy reward converges rapidly and plateaus earlier than the 488-sample run. In addition, the format reward follows a similar convergence pattern across all sample sizes, likely reflecting the robust structural adherence capability of the Qwen2.5-VL instruct model.

We extensively evaluate the resulting model checkpoints across three chart-related benchmarks.

Category	Fine-tuning strategy	
	SFT	Chart-RL
annotations	<u>0.41</u>	<u>0.41</u>
benchmark	0.47	0.52
color_random	<u>0.37</u>	<u>0.37</u>
color_scheme	<u>0.34</u>	<u>0.34</u>
data_pivot	0.35	0.36
font	0.39	0.41
grid	0.38	0.41
hatching	0.28	0.34
horizontal_grouped	0.35	0.38
horizontal_stacked	0.34	0.37
legend_position	0.34	0.40
line_representation	0.32	0.35
log_scale	0.33	0.30
normal	0.37	0.41
only_data_color		
scheme	<u>0.35</u>	<u>0.35</u>
replacing_legend		
with_labels	0.35	0.36
scaling	0.34	0.32
scatter_representation	<u>0.31</u>	<u>0.31</u>
stacked	0.30	0.32
stacked_area	0.32	0.33
stair_plot_normal	0.23	0.29
stair_plot_with_marker	0.32	0.35
stem_plot	0.30	0.31
tick_orientation	0.37	0.43
tick_position	0.36	0.37

Table 3: Robustness comparison between Chart-RL and SFT across 25 perturbed chart categories.

As shown in Table 4, Chart-RL produces substantial improvements on MultiChartQA and ChartInsights despite using far fewer training samples.

In summary, Chart-RL demonstrates strong generalization across diverse benchmarks while requiring only minimal training examples. Unlike traditional SFT, this data-efficient training paradigm enables effective adaptation to both novel chart tasks and predefined output formats.

4.5 Out-of-Domain Generalization

We investigate whether Chart-RL’s reasoning capabilities generalize to out-of-domain visual mathematical tasks. As shown in Table 5, Chart-RL achieves a significant 55.6% relative improvement on MathVerse (Zhang et al., 2024), despite not being explicitly trained on mathematical reasoning data. These results demonstrate that Chart-RL effectively transfers to visually grounded mathe-

Method	Training Task	Training Samples	MultiChartQA			ChartInsights		
			Accuracy	Rel. Δ	p -value	Accuracy	Rel. Δ	p -value
Baseline	–	–	35.2	–	–	36.4	–	–
Chart-RL (ours)	Hard	10	39.1	+11.1%	$< 10^{-2}$	39.1	+7.4%	$< 10^{-4}$
		100	<u>39.2</u>	+11.4%	$< 10^{-2}$	<u>39.7</u>	+9.1%	$< 10^{-6}$
		448	41.1	+16.7%	$< 10^{-4}$	40.6	+11.5%	$< 10^{-6}$

Table 4: Evaluation results for Chart-RL across different training sample sizes, showing performance on three chart understanding benchmarks under varying RL training data scales. The rightmost column for each dataset shows the proportional test results for the relative change. All reported results are statistically significant at $\alpha = 0.05$.

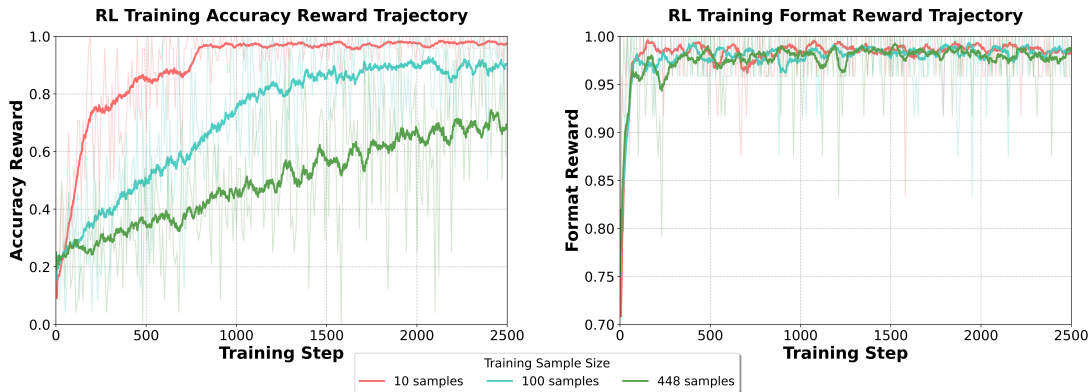


Figure 3: RL Training reward trajectories for Chart-RL across different training sample sizes, showing accuracy rewards (left) and format rewards (right) throughout the training process.

mathematical problem solving.

Our findings align with recent studies highlighting domain-specific reinforcement learning can incentivize transferable analytical skills (Xie et al., 2025). We believe that complex chart-based reasoning represents an important yet underexplored training data source that could benefit the development of generalized VLMs (Liang et al., 2024).

Methods	MathVerse		
	Vision		
	Acc.	Rel. Δ	p -value
Baseline	1.8	–	–
Chart-RL (ours)	2.8	+55.6%	$< 10^{-2}$

Table 5: Evaluation results on MathVerse Vision.

5 Ablation Study

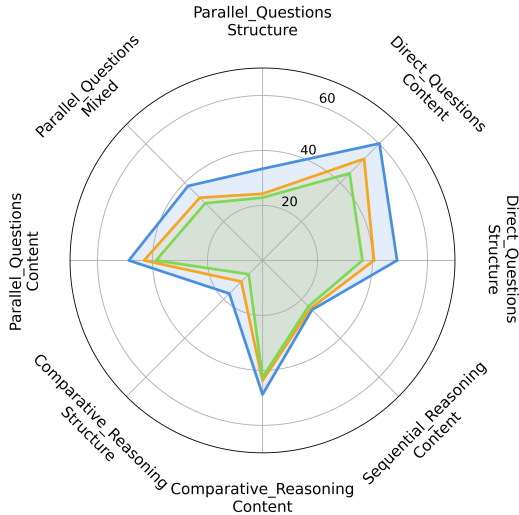
5.1 Generalization in Low-Level Tasks

To evaluate Chart-RL’s generalization capabilities within the chart understanding domain, we compare Chart-RL against the SFT-only variant and

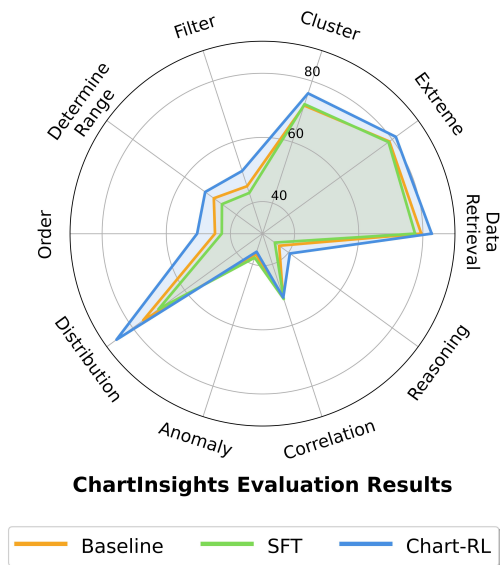
the base Qwen2.5-VL-3B-Instruct model. Figure 4 presents detailed per-category performance comparisons on ChartInsights and MultiChartQA benchmarks.

Chart-RL consistently outperforms both the baseline and the SFT-enhanced variant across various low-level tasks. Our method demonstrates robust improvements, likely resulting from the reinforced exploration on challenging chart-query pairs during RL training. Specifically, Chart-RL achieves substantial gains in reasoning tasks (Direct Questions, Parallel Questions, and Comparative Reasoning) on MultiChartQA, and data operations (Order, Filter, and Cluster) on ChartInsights.

By training on highly complex chart samples, we enable effective transfer to simpler, low-level chart-related tasks without requiring additional task-specific training. This suggests that our method facilitates fundamental chart understanding capabilities that generalize beyond the predefined training tasks.



MultiChartQA Evaluation Results



ChartInsights Evaluation Results

Figure 4: Sub-category analysis of Qwen-2.5VL-3B performance on the MultiChartQA and ChartInsights benchmarks.

5.2 Task Complexity Matters: Easy vs. Hard Training Tasks

To investigate how the complexity of training tasks affects RL performance, we compare models trained on tasks of varying difficulty. Figure 5 shows drastically different learning dynamics. Specifically, easy tasks achieve high accuracy rapidly but offer limited learning potential, plateauing around 0.9 with minimal subsequent improvement. In contrast, hard tasks begin with lower accuracy of 0.2 but demonstrate sustained improvement throughout RL training. This continuous learning trajectory suggests that hard tasks provide helpful feedback signals, enabling the model to develop

more sophisticated chart reasoning capabilities.

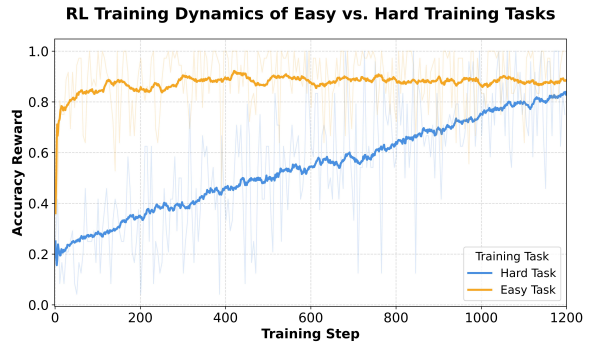


Figure 5: Accuracy reward trajectories in RL training for Easy and Hard tasks.

Table 6 shows the impact of task complexity in RL training. These results highlight a critical insight: training on challenging tasks that maintain a consistent learning signal throughout optimization leads to better generalization. The continuous improvement on hard tasks suggests that the model develops robust chart comprehension capabilities, while the early saturation on easy tasks likely leads to overfitting on limited patterns. Our results indicate that the complexity and diversity of training tasks are essential in RL for developing generalizable chart comprehension capabilities.

Task	MultiChartQA	ChartInsights
Baseline	35.2	36.4
Easy	34.8 (-1.2%)	36.0 (-1.1%)
Hard	41.1 (+16.8%)	40.6 (+11.5%)

Table 6: Evaluation results of Easy vs. Hard training tasks on different benchmarks.

6 Conclusion

We present Chart-RL, an effective RL-based method that significantly advances task-agnostic chart comprehension capabilities in VLMs. Through systematic experiments across diverse benchmarks, we demonstrate that Chart-RL achieves substantial improvements using only a few hundreds of training samples. Our analysis reveals that training on complex tasks that provide continuous learning signals produces better generalization compared to optimizing for high training accuracy on simpler tasks. Moreover, Chart-RL exhibits cross-domain transfer capability, improving multimodal mathematical reasoning performance without explicit training in this domain.

Limitations

Our work has a few limitations that should be acknowledged. First, Chart-RL relies on mathematically verifiable ground truths for reward computation, which restricts its applicability to chart understanding tasks that lack deterministic answers or require subjective interpretation. Second, the evaluation remains relatively narrow and may not capture the full diversity of real-world chart comprehension scenarios. The Chart-RL training requires careful curation of complex chart reasoning tasks, which may not be readily available for all domains or chart types. Third, our work mainly focus on the initial application of RLVR for chart question answering. We believe that a multi-stage post-training strategy, alternating SFT and RL, might lead to enhanced accuracy and generalization.

References

- Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. 2025. Small language models are the future of agentic ai. *arXiv preprint arXiv:2506.02153*.
- Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, and 1 others. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.
- 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. *GitHub*. Accessed: 2025-02-02.
- Huilin Deng, Ding Zou, Rui Ma, Hongchen Luo, Yang Cao, and Yu Kang. 2025. Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning. *arXiv preprint arXiv:2503.07065*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Y. Han, C. Zhang, X. Chen, X. Yang, Z. Wang, G. Yu, B. Fu, and H. Zhang. 2023. ChartLlama: a multi-modal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Kung-Hsiang Huang, Hou Pong Chan, Yi R Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. 2024. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models. *IEEE Transactions on Knowledge and Data Engineering*.
- Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi R Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. 2023. Do llms understand charts? analyzing and correcting factual errors in chart captioning. *arXiv preprint arXiv:2312.10160*.
- Muye Huang, Han Lai, Xinyu Zhang, Wenjun Wu, Jie Ma, Lingling Zhang, and Jun Liu. 2025a. Evochart: A benchmark and a self-training approach towards real-world chart understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3680–3688.
- Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Yao Hu, and Shaohui Lin. 2025b. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.
- Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.
- Damjan Kalajdziewski. 2024. Scaling laws for forgetting when fine-tuning large language models. *arXiv preprint arXiv:2401.05605*.
- Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's "up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*.
- G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*.
- K. Lee, M. Joshi, I. Turc, H. Hu, F. Liu, J. Eisen-schlos, U. Khandelwal, P. Shaw, M.W. Chang, and K. Toutanova. 2023. Pix2struct: screenshot parsing as pretraining for visual language understanding. In *Proceedings of the 40th International Conference on Machine Learning, ICML'23*. JMLR.org.
- Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. 2024. A survey of multimodal large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409.

- Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhui Chen, Nigel Collier, and Yasemin Altun. 2022. Deplot: One-shot visual language reasoning by plot-to-table translation. *arXiv preprint arXiv:2212.10505*.
- H. Liu, C. Li, Q. Wu, and Y.J. Lee. 2023. Visual instruction tuning. In *NeurIPS*.
- Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.
- A. Masry, P. Kavehzadeh, X.L. Do, E. Hoque, and S. Joty. 2023. UniChart: a universal vision-language pretrained model for chart comprehension and reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14662–14684, Singapore. Association for Computational Linguistics.
- Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.
- Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhui Wang, Junjun He, Kaipeng Zhang, and 1 others. 2025. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*.
- Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536.
- Srija Mukhopadhyay, Adnan Qidwai, Aparna Garimella, Pritika Ramu, Vivek Gupta, and Dan Roth. 2024. Unraveling the truth: Do vlms really understand charts? a deep dive into consistency and robustness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16696–16717.
- Shubham Parashar, Shurui Gui, Xiner Li, Hongyi Ling, Sushil Vemuri, Blake Olson, Eric Li, Yu Zhang, James Caverlee, Dileep Kalathil, and 1 others. 2025. Curriculum reinforcement learning from easy to hard tasks improves llm reasoning. *arXiv preprint arXiv:2506.06632*.
- Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. 2025. Lmm-rl: Empowering 3b llms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*.
- Florian Schleid, Jan Strich, and Chris Biemann. 2025. Visual question answering on scientific charts using fine-tuned vision-language models. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 211–220.
- Luca M. Schulze Buschoff, Konstantinos Voudouris, Elif Akata, Matthias Bethge, Joshua B. Tenenbaum, and Eric Schulz. 2025. Testing the limits of fine-tuning for improving visual cognition in vision language models. In *International Conference on Machine Learning*. Published as a poster.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseek-math: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, and 1 others. 2025. Vlm-rl: A stable and generalizable rl-style large vision-language model. *arXiv preprint arXiv:2504.07615*.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, and 1 others. 2025. Reinforcement learning for reasoning in large language models with one training example. *arXiv preprint arXiv:2504.20571*.
- Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, and 1 others. 2024. Chartxv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697.
- Yifan Wu, Lutao Yan, Leixian Shen, Yunhai Wang, Nan Tang, and Yuyu Luo. 2024. Chartinsights: Evaluating multimodal large language models for low-level chart question answering. *arXiv preprint arXiv:2405.07001*.
- Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*.
- Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, and 1 others. 2025. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.

Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. 2025. R1-zero's "aha moment" in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*.

Zifeng Zhu, Mengzhao Jia, Zhihan Zhang, Lang Li, and Meng Jiang. 2024. Multichartqa: Benchmarking vision-language models on multi-chart problems. *arXiv preprint arXiv:2410.14179*.

A Appendix

A.1 CoT Prompt Template

We use the GPT-4o model via Azure OpenAI (API version 2024-12-01-preview) to generate explicit chain-of-thought (CoT) reasoning. Below is the prompt template:

Prompt for Generating CoT Trajectories with GPT-4o

I will give you an image, an original question, and its correct answer. Your task is to rewrite the question so that solving it requires step-by-step CoT reasoning, including any necessary numerical or mathematical expressions. Feel free to include natural thought markers (e.g., ‘let me think,’ ‘oh, I see’).

- The rewritten question must request a specific, easily verifiable answer (for example, ‘2’ or ‘A’)—no open-ended prompts.
- Do not include the word ‘Answer:’ in the question to avoid leakage.

Input: Original Question: {original_question}
Original Answer: {original_answer}

Output: Question: [Rewritten question] Answer: [Detailed reasoning steps, with calculations as needed] <think>Step-by-step reasoning process</think> <answer>Final, verifiable answer</answer>

A.2 Examples of Easy and Hard Tasks

Easy tasks typically involve direct information extraction where the answer can be found through simple visual inspection, such as reading a single value from a clearly labeled data point. In contrast, hard tasks require multi-step reasoning, including comparing multiple data points, identifying relationships between variables, and performing intermediate calculations or rankings.

The query for the easy task in Figure 6 is: “What is the sodium content of sample 34?” The corresponding ground truth annotation is 12.8.

The query for the hard task in Figure 7 is: “Provide the accuracy (%) of the model with the second highest time.” The corresponding ground truth annotation is 94.

The corresponding chain-of-thought (CoT) reasoning trajectory from GPT-4o is shown below:

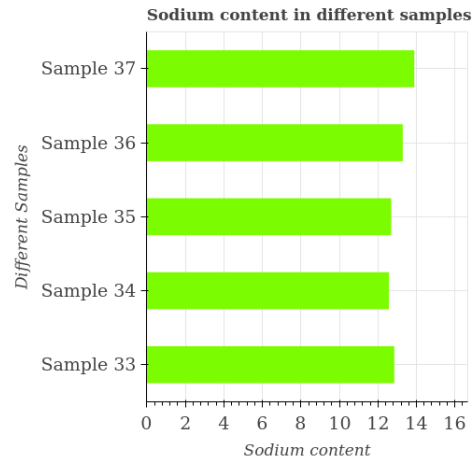


Figure 6: An example of easy task for RL training.

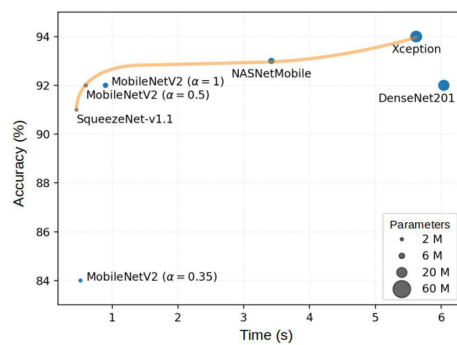


Figure 7: An example of hard task for RL training.

CoT Reasoning Trajectory for the Hard Example

<think>To determine the accuracy of the model with the second highest execution time, we **first** observe the x-axis labeled “Time (s)” and identify the models based on their execution times. The longest time corresponds to the model Xception at approximately 6 seconds. The second longest time corresponds to DenseNet201 at approximately 5 seconds. **Next**, we look at the y-axis labeled “Accuracy (%)” to find the accuracy value for DenseNet201. The accuracy for DenseNet201 is clearly marked as 94%.</think>

<answer>94</answer>