CHART-SR: CHART SENSITIVE REASONING WITH VERIFIABLE REWARDS IN MULTIMODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Chart understanding with multimodal large language models (MLLMs) has advanced rapidly, yet existing methods mainly supervise final answers or reasoning outcomes, providing limited constraints on whether predictions are grounded in fine-grained, question-relevant visual evidence. This limitation is particularly critical for charts, where small errors in values, scales, or visual mappings can lead to substantial answer errors. To address this issue, we propose Chart Sensitive Reasoning (Chart-SR), a framework that leverages the structured nature of charts for controlled interventions, transforming visual-evidence dependence into verifiable supervision. We further construct the Chart-Sensitivity Dataset (CS-Dataset), a programmatically generated and fully verifiable dataset of paired charts. Given an original chart, an intervened chart, and a shared question, Chart-SR learns to answer the original chart, identify visual changes, judge their effects on the answer, and answer the intervened chart accordingly. We further optimize this structured reasoning process with GRPO using multi-component verifiable rewards for answer correctness, change identification, and answer-effect judgment. Extensive experiments across multiple benchmarks show that Chart-SR consistently improves chart reasoning and enhances sensitivity to fine-grained visual evidence.

1 INTRODUCTION

Multimodal large language models (MLLMs) have achieved substantial progress in visual question answering, understanding, and complex reasoning Amirloo et al. (2024); Liu et al. (2024b); Kuang et al. (2025), and have increasingly been applied to real-world scenarios such as medical data analysis Chen et al. (2024a) and financial decision-making Liu et al. (2024a). Charts are highly structured visual representations that compactly encode numerical values, categories, trends, and their relationships, serving as an important medium for communicating quantitative information Zhang et al. (2026); Suri et al. (2025). Unlike natural-image understanding, which mainly relies on object and scene semantics, chart understanding requires jointly interpreting textual labels, visual marks, coordinate scales, legend mappings, and numerical relations, followed by information retrieval, comparison, calculation, and multi-step reasoning Huang et al. (2026). Because charts are information-dense, chart reasoning is particularly sensitive to fine-grained visual evidence: even a small error in reading a value, scale, or visual mapping can be amplified into a substantial numerical error . Therefore, reliable chart reasoning requires not only correct predictions, but also accurate grounding in the fine-grained visual evidence that supports them Masson et al. (2023); Akhtar et al. (2024).

Recent advances in chart understanding have substantially improved MLLMs' ability to extract visual information, perform numerical reasoning, and generate structured responses Xu et al. (2023); Masry et al. (2025). Existing approaches mainly enhance chart reasoning through chart-specific instruction tuning, structured intermediate representations, and reinforcement learning (RL) Schulman et al. (2017) with verifiable objectives Masry et al. (2024); Liu et al. (2023). These efforts have advanced chart understanding from basic information extraction toward structured, multi-step, and verifiable reasoning Akhtar et al. (2023); Chen et al. (2026); Suri et al. (2025); Vogel et al. (2025). However, existing supervision still primarily focuses on final answers, reconstructed structures, or reasoning outcomes within individual charts. It provides limited explicit guidance on whether models correctly identify and rely on the fine-grained visual evidence required by the question. Consequently, even a correct answer does not necessarily reveal whether the model has truly grounded its prediction in the decisive values, labels, or visual mappings.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

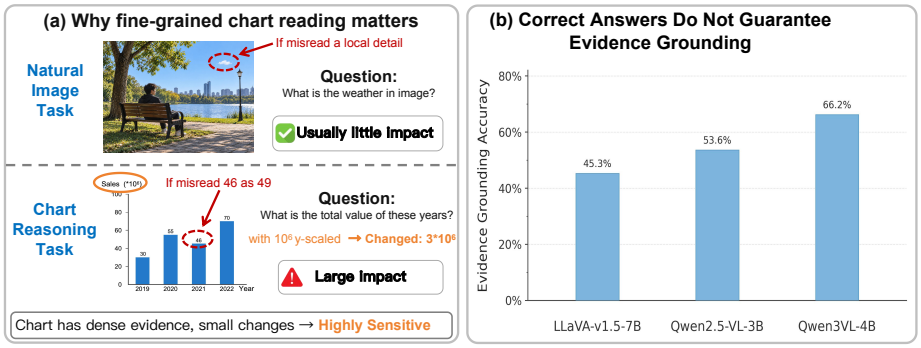


Figure 1: Motivating examples of our work. (a) Charts encode dense quantitative evidence, where even subtle visual errors may affect answer-critical information. (b) Statistical experiments with LLaVA-v1.5-7B, Qwen2.5-VL-3B and Qwen3-VL-4B on the PlotQA dataset Methani et al. (2020), demonstrating a gap between answer correctness and evidence grounding accuracy.

This limitation is particularly critical for charts. As illustrated in Fig. 1(a), many natural-image questions can still be resolved from coarse object or scene-level semantics despite minor local visual errors. In contrast, charts compress dense quantitative information into limited visual space, where a seemingly small perception error, such as interpreting 46 as 49 under a 10^6-scaled y-axis, can lead to a substantial numerical discrepancy. Chart reasoning therefore often depends on precise, fine-grained visual evidence, whereas conventional answer- or instance-level supervision only verifies whether the final prediction is correct. A model may still arrive at the correct answer through question templates, dataset biases, linguistic priors, or parametric memory, without faithfully grounding its prediction in the chart. As shown in Fig. 1(b), MLLMs exhibit limited evidence grounding even on correctly answered PlotQA samples, showing that answer correctness does not guarantee faithful grounding. Even structured reasoning or grounding supervision does not directly verify whether the model's prediction truly depends on the evidence required by the question. This reveals a fundamental supervision-granularity mismatch: chart reasoning requires fine-grained visual evidence, while existing training signals remain largely outcome-level and provide limited supervision over the underlying evidence dependence.

Therefore, an important question in chart reasoning is how to transform otherwise implicit visual-evidence dependence into fine-grained, verifiable supervision, such that models not only answer correctly but also identify and faithfully rely on the evidence required by the question. To address this problem, we propose Chart Sensitive Reasoning (Chart-SR), a framework for learning evidence-sensitive chart reasoning with verifiable reinforcement learning. Our key idea is to exploit the programmatically editable nature of charts and construct minimal controlled interventions over values, labels, or visual attributes while keeping the remaining content unchanged. Rather than treating chart perturbation as a standalone application task, these controlled interventions serve as a supervision mechanism that exposes how model predictions depend on visual evidence. Built upon the newly constructed Chart-Sensitivity Dataset (CS-Dataset), Chart-SR explicitly learns to answer the original chart, identify the visual change, determine its effect on the current answer, and produce the corresponding answer for the intervened chart. We further optimize this structured reasoning process with Group Relative Policy Optimization (GRPO) using verifiable rewards for answer correctness, change identification, and answer-effect judgment, encouraging sensitivity to relevant evidence while preserving invariance to irrelevant changes. Extensive experiments demonstrate that Chart-SR yields notable gains in evidence grounding, chart understanding, and reliable reasoning.

Our main contributions are summarized as follows:

1. We propose Chart Sensitive Reasoning (Chart-SR), a framework for fine-grained visual evidence-aware reasoning. Chart-SR extends answer-level supervision into a structured reasoning involving original-chart answering, visual-change identification, answer-effect judgment, and intervened-chart answering, explicitly modeling the dependence between predictions and visual evidence.

2. We introduce Chart-Sensitivity Dataset (CS-Dataset), a fully verifiable dataset of controlled chart interventions that provides fine-grained supervision for visual-evidence learning.

3. We develop a GRPO-based training strategy with multi-component verifiable rewards. The rewards jointly supervise answer correctness, visual-change identification, and answer-effect judgment, enabling fine-grained visual perception and evidence-sensitive reasoning.

4. We conduct extensive experiments across multiple benchmarks. Chart-SR consistently improves general chart reasoning and substantially enhances sensitivity to fine-grained visual evidence.

2 Related Works

2.1 Multimodal Reasoning with Verifiable RL

Recent advances in multimodal reasoning have evolved from general reasoning-oriented RL paradigms to evidence-grounded and causally consistent reasoning. OpenAI o1 Jaech et al. (2024) and DeepSeek-R1 Shao et al. (2024); Guo et al. (2025) demonstrated the effectiveness of large-scale reinforcement learning (RL) for eliciting long-chain reasoning, while GRPO enabled scalable optimization without an explicit critic model. Vision-R1 Huang et al. (2025b) and R1-VL Zhang et al. (2025) further extended RL-based reasoning to MLLMs through multimodal CoT cold-start initialization and step-wise reward optimization, respectively. However, these methods mainly optimize instance-level correctness and format validity. UV-CoT Zhao et al. (2025) and GRIT Fan et al. (2026), introduce explicit visual grounding into reasoning processes, advancing from text-only CoT toward visually grounded reasoning. More recently, DeFacto Xu et al. (2025a) and CFPO Yu et al. (2026) leverage counterfactual supervision and policy optimization to enhance the performance.

2.2 Chart Understanding and Reasoning

Chart understanding has evolved from representation learning to visual grounding, complex reasoning and RL. Early works, including DePlot Liu et al. (2023) and ChartVLM Xia et al. (2025), established key paradigms through plot-to-table conversion, and chart-specific modeling. Subsequent works such as ChartGemma Masry et al. (2025), ChartLens Suri et al. (2025), and RefChartQA Vogel et al. (2025) further improved direct chart-image understanding, fine-grained visual attribution, and evidence-grounded answering. Recent efforts have explored more challenging scenarios: ReFocus Fu et al. (2025), ChartSketcher Huang et al. (2026), ChartPoint Xu et al. (2025b), ChartFR1 Pan et al. (2026), and VisDoT Lee et al. (2026) enhanced interpretable reasoning via visual editing, sketch-based feedback, grounding reflection, focus-driven RL, and perception–logic decomposition. Moreover, Chart-R1 Chen et al. (2026) incorporated CoT supervision and numerical RL, while MSRL Chen et al. (2025) developed structured RL for chart-to-code generation.

3 Methodology

Figure 2 illustrates the overall framework. We first formulate Chart-SR reasoning task (Section 3.1). Then, we construct verifiable chart pairs with perturbation annotations (Section 3.2). Finally, we introduce the GRPO training with multi-component verifiable rewards (Section 3.3).

3.1 Problem Formulation, and Chart-SR Structured Reasoning Chain

Problem Formulation: Let a chart be generated from its annotation A through a rendering function $\mathcal{R}(\cdot)$, where A contains numerical values, data series, axes, legends, and other chart elements. Given a controlled perturbation operator δ, we define the original chart as $I_1 = \mathcal{R}(A)$ and the perturbed chart as $I_2 = \mathcal{R}(\delta(A))$.

The perturbation is applied to the structured annotation rather than directly to the image pixels. Moreover, both charts share the same rendering configuration. Consequently, (I_1, I_2) forms a minimal chart pair whose differences are restricted to the designated perturbation.

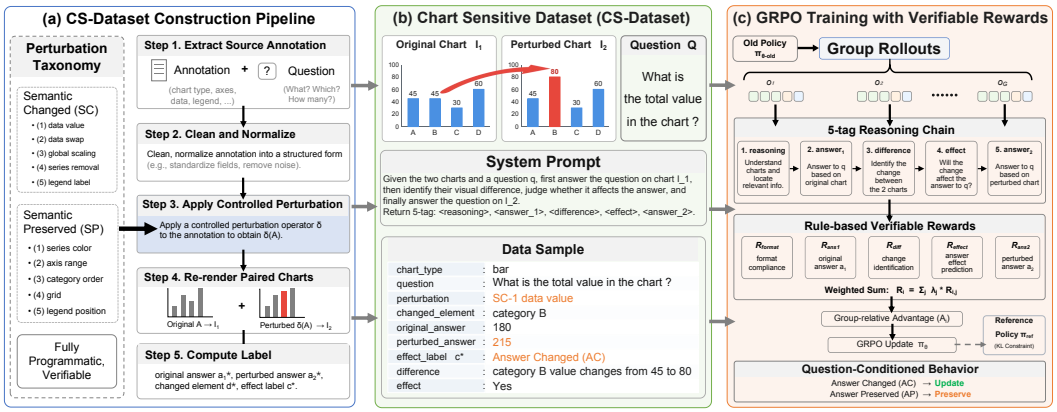


Figure 2: Framework of Chart-SR. We first construct Chart Sensitive Dataset (CS-Dataset) with verified answers and perturbation effects. Given an original chart, a perturbed chart, and a shared question, Chart-SR generates a structured five-tag reasoning trajectory consisting of reasoning over the charts, original answer prediction, difference identification, answer-effect judgment, and perturbed answer prediction. We optimize the model with GRPO using verifiable rewards for answer correctness, change identification, and answer-effect prediction.

We define the question-conditioned perturbation label as

$$e^\star = \begin{cases} \text{Answer} - \text{Changed}(\mathbf{AC}), & a_1^\star \neq a_2^\star, \\ \text{Answer} - \text{Preserved}(\mathbf{AP}), & a_1^\star = a_2^\star, \end{cases} \tag{1}$$

Given the paired charts and their shared question, the model policy π_θ generates:

$$(\hat{a}_1, \hat{d}, \hat{e}, \hat{a}_2) \sim \pi_\theta(\cdot \mid I_1, I_2, q), \tag{2}$$

where $\hat{a}_1$ and $\hat{a}_2$ are the predicted answers for the original and perturbed charts, respectively; $\hat{d}$ describes the perturbation; and $\hat{e} \in \text{AC}, \text{AP}$ predicts whether the perturbation changes the answer.

Unlike conventional image-difference detection, our task does not aim to identify visual changes alone. The effect of a perturbation is jointly determined by δ and q: the model must decide whether the change affects question-relevant evidence and thus requires answer updating. Therefore, predicting $\hat{a}_2$ is not equivalent to independently answering two chart questions, but requires identifying the intervention, judging its effect, and updating or preserving the answer accordingly.

Structured Reasoning Chain: To make the reasoning process explicitly supervisable and verifiable, we organize the model output into the following 5-tag: ⟨reasoning⟩ for reasoning over both charts, ⟨answer_1⟩ for the original answer, ⟨difference⟩ for identifying chart changes, ⟨effect⟩ for judging whether the intervention affects the answer, and ⟨answer_2⟩ for the perturbed answer. Accordingly, the reasoning trajectory follows reasoning → answer$_1$ → difference → effect → answer$_2$. This chain explicitly separates chart-change perception from effect judgment, while providing verifiable intermediate signals for optimizing and evaluating each reasoning step.

3.2 Perturbed Chart-Pair Construction

3.2.1 Data Source and Re-rendering Pipeline

We construct the training data from PlotQA Methani et al. (2020), which comprises 28.9M Chart QA pairs grounded over 224,377 real-world plots. The construction pipeline consists of 5-step. We first clean and normalize the original chart annotations. We then apply a perturbation operator δ to the structured annotation and obtain $\delta(A)$. Next, A and $\delta(A)$ are rendered under identical parameters to produce I_1 and I_2. The same question q is assigned to both charts. Finally, we compute $a_1^\star$ and $a_2^\star$.

All annotations are generated by programs rather than human annotators or LLMs, making them reproducible and auditable. Our charts cover four types: vertical bar, horizontal bar, line, dot-line.

Table 1: Taxonomy of controlled chart interventions. SC denotes Semantics-changed perturbations, while SP denotes Semantics-preserved perturbations. Details can be found in Appendix A.1.

Type	Perturbation 1	Perturbation 2	Perturbation 3	Perturbation 4	Perturbation 5
SC	SC-1 Data value	SC-2 Data swap	SC-3 Global scaling	SC-4 Series removal	SC-5 Legend label
SP	SP-1 Series color	SP-2 Axis range	SP-3 Category order	SP-4 Grid	SP-5 Legend position

3.2.2 PERTURBATION TAXONOMY

We divide the perturbations into two categories, targeting invariance and sensitivity. Table 1 summarizes the two categories of controlled chart interventions and their corresponding types.

Semantics-changed (SC) perturbations. The SC perturbations modify the data, structure, or label semantics of a chart. They include changing a data-point value (SC-1), swapping two data points (SC-2), proportionally scaling all values (SC-3), removing a data series (SC-4), and modifying a legend label (SC-5). Whether an SC perturbation results in AC or AP depends on whether the modified element is relevant to the current question.

Semantics-preserved (SP) perturbations. The SP perturbations modify only the visual presentation of a chart without changing its underlying data semantics. They include series-color modification (SP-1), axis-range adjustment (SP-2), category-order rearrangement (SP-3), grid-line modification (SP-4), and legend-position adjustment (SP-5). For any question q,

$$f(A, q) = f\big(\delta(A), q\big), \qquad \delta \in \Delta_{\mathrm{SP}}, \tag{3}$$

and all SP samples are therefore labeled as AP. These samples teach the model to remain invariant to visually altered but semantically irrelevant evidence.

The SC-AP subset is particularly important, because the chart semantics have genuinely changed, but the change is irrelevant to the current question. The model must perceive the perturbation, determine its irrelevance, and preserve the answer. Details can be found in Appendix A.1.

3.2.3 PROGRAMMATIC ANSWER COMPUTATION AND LABEL BALANCING

For each chart pair, $a_1^\star$, $a_2^\star$, and $d^\star$ are directly computed from the structured annotations. A second-pass verification confirms full consistency between the answers and the derived perturbation labels.

To prevent the model from exploiting correlations between perturbation categories and effect labels, we rebalance the AC/AP ratio to exactly 1:1. The complete RL dataset has an approximate AC/AP ratio of 50:50. The questions cover arithmetic operations (39.7%), comparisons (28.3%), extrema (19.4%), and value retrieval (12.6%). The four chart types are approximately balanced.

3.3 TRAINING WITH VERIFIABLE REWARDS

3.3.1 MULTI-COMPONENT VERIFIABLE REWARDS

We employ five types of verifiable reward signals, covering paired-chart numerical answers, perturbation-effect judgment, perturbation identification, and output format. Since all ground-truth answers are numerical, we uniformly use a relative tolerance of $\tau_{\mathrm{num}} = 0.05$ for answer verification. Given a parsed prediction $\hat{a}_k$ and its ground-truth answer $a_k^\star$, where $k \in \{1, 2\}$ denotes the original and perturbed charts, respectively, we define the answer-correctness indicator as

$$c_k = \begin{cases} \mathbb{I}[|\hat{a}_k - a_k^\star| \le \tau_{\mathrm{num}} |a_k^\star| + \epsilon_{\mathrm{num}}], & a_k^\star \neq 0, \\ \mathbb{I}[\hat{a}_k = 0], & a_k^\star = 0. \end{cases} \tag{4}$$

Accordingly, we set $R_{\mathrm{ans1}} = c_1$ and $R_{\mathrm{ans2}} = c_2$. Let $d^\star$ be the ground-truth perturbed elements, and $\mathrm{parse}(\hat{d})$ extract them from the generated difference description. The rewards are: $R_{\mathrm{diff}} = \mathbb{I}[\mathrm{parse}(\hat{d}) = d^\star]$, $R_{\mathrm{effect}} = \mathbb{I}[\mathrm{norm}_e(\hat{e}) = e^\star]$. The format reward R_{format} verifies whether all five required tags are present, correctly ordered, and successfully parsed. The total reward is

$$R = 0.05 R_{\mathrm{format}} + 0.25 R_{\mathrm{ans1}} + 0.30 R_{\mathrm{ans2}} + 0.225 R_{\mathrm{diff}} + 0.175 R_{\mathrm{effect}}. \tag{5}$$

All reward components are deterministic and rule-based, requiring no external LLM judge and ensuring reproducible, low-cost optimization. The weights prioritize answer correctness to prevent intermediate-reward exploitation from degrading task accuracy, while retaining substantial supervision for difference identification and perturbation-effect judgment. Thus, Chart-SR penalizes failures in visual-evidence reasoning even when the final answer is accidentally correct.

3.3.2 GROUP RELATIVE POLICY OPTIMIZATION (GRPO)

GRPO estimates advantages by comparing the rewards of multiple responses sampled for the same input. For each input $x = (I_1, I_2, q)$, we sample a group of G responses from the old policy:

$$o_1, \ldots, o_G \sim \pi_{\theta_{\text{old}}}(\cdot \mid x), \tag{6}$$

and compute their rewards $r_1, \ldots, r_G$ using Eq. 5. Then the group-relative advantage is

$$\widehat{A}_i = \frac{r_i - \mu_r}{\sigma_r + \epsilon_A}, \qquad \mu_r = \frac{1}{G}\sum_{j=1}^{G} r_j, \tag{7}$$

where σ_r is the standard deviation of the group rewards and $\epsilon_A > 0$ is a numerical stabilizer. The outcome-level advantage $\widehat{A}_i$ is assigned to every token in response o_i.

For the t-th token in o_i, the importance-sampling ratio is

$$\rho_{i,t}(\theta) = \pi_\theta(o_{i,t} \mid x, o_{i,<t}) \,/\, \pi_{\theta_{\text{old}}}(o_{i,t} \mid x, o_{i,<t}). \tag{8}$$

Let $\widehat{A}_i$ be the group-relative advantage of o_i, and let $D_{\text{KL}}^{i,t}$ denote the token-level KL penalty. Define the per-token clipped objective as

$$\ell_{i,t}(\theta) = \min\Big[\rho_{i,t}(\theta)\widehat{A}_i,\ \text{clip}(\rho_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon)\,\widehat{A}_i\Big] - \beta D_{\text{KL}}^{i,t}. \tag{9}$$

Then the GRPO objective is defined as

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\ell_{i,t}(\theta)\right], \tag{10}$$

where ε controls the policy-update range and β controls the KL regularization strength.

Following GRPO, we define the token-level log-probability ratio as $r_{i,t} = \log \pi_\theta(o_{i,t} \mid x, o_{i,<t}) - \log \pi_{\text{ref}}(o_{i,t} \mid x, o_{i,<t})$, and estimate the token-level KL divergence as $D_{\text{KL}}^{i,t} = \exp(-r_{i,t}) + r_{i,t} - 1$.

4 EXPERIMENTS

4.1 DATASETS AND EVALUATION METRICS

Training Data. We construct our training data from the training split of PlotQA Methani et al. (2020). Our training dataset contains 6,048 RL samples, with 3,024 AC and 3,024 AP samples. To prevent evaluation leakage, all of our training data are constructed from the PlotQA training split, while PlotQA evaluation follows the standard protocol by using the corresponding test split.

Evaluation Benchmarks. We evaluate general chart understanding on ChartQA Masry et al. (2022), PlotQA Methani et al. (2020), and Evo-ChartQA Huang et al. (2025a) for chart information extraction and question answering, as well as ChartBench Xu et al. (2023) and Overall visual evaluation for complex visual reasoning over charts.

To better assess training effectiveness, we construct Chart-Perturb as a strictly disjoint diagnostic split of CS-Dataset. Chart-Perturb mainly with two perspectives: intervention awareness and effect consistency. We report All Accuracy (Acc_{All}), AC Accuracy (Acc_{AC}), and AP Accuracy (Acc_{AP}). AccAll measures the overall accuracy of perturbation-effect prediction across AC and AP samples:

$$\text{Acc}_{All} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(\hat{e}_i = e_i^\star\right). \tag{11}$$

The AC/AP accuracy evaluates whether the model correctly identifies AC and AP interventions:

$$\text{Acc}_{AC} = \frac{1}{N_{AC}}\sum_{i\in AC}\mathbb{I}(\hat{e}_i = AC), \quad \text{Acc}_{AP} = \frac{1}{N_{AP}}\sum_{i\in AP}\mathbb{I}(\hat{e}_i = AP). \tag{12}$$

Table 2: Comparison results on chart understanding benchmarks, including ChartQA Masry et al. (2022), EvoChart-QA Huang et al. (2025a), ChartBench Xu et al. (2023), PlotQA Methani et al. (2020) and Overall Visual results. The best results are highlighted in **bold**.

Model	ChartQA-H	ChartQA-A	EvoChart-QA	ChartBench	PlotQA	Overall
Proprietary models						
GPT-4o	84.32	88.48	52.80	61.47	42.96	66.01
Gemini-2.0	84.00	88.24	**64.64**	55.63	63.36	71.17
Claude-3.5	85.04	90.72	56.96	56.96	57.63	69.46
Open-source models and chart models						
Qwen2VL-2B	50.48	75.84	23.84	20.27	38.80	41.85
Qwen2VL-72B	82.48	88.56	54.00	54.77	73.76	70.71
Qwen2.5VL-7B	78.76	93.76	53.50	66.40	55.50	69.58
QvQ-Preview-72B	83.20	89.76	54.32	42.40	69.04	67.74
InternVL2.5-78B	78.48	89.44	57.44	65.57	57.20	69.63
ChartGemma-2B	53.44	86.64	36.08	23.87	25.76	45.16
ChartSketcher-2B	55.60	80.88	26.72	30.10	41.12	46.88
ChartSketcher-72B	**85.20**	92.64	63.28	68.33	76.72	77.23
Chart-Point-7B	81.36	94.12	-	63.04	–	–
Our models						
Chart-SR-4B(Ours)	78.00	94.45	60.70	68.60	78.20	75.99
Chart-SR-8B(Ours)	79.20	**94.70**	63.90	**69.20**	**80.30**	**77.46**

Table 3: Results on chart reasoning benchmarks, including MultiChart-QA Zhu et al. (2025), RobustCQA Mukhopadhyay et al. (2024) and Chart-Perturb. The best results are highlighted in **bold**.

Method	Params.	Chart-Perturb			MultiChart-QA			RobustCQA		
		Acc_{All}	Acc_{AC}	Acc_{AP}	Overall	Sequential	Comparative	Overall	Augmented	Human
Qwen3-VL Bai et al. (2025a)	4B	0.56	0.47	0.66	66.35	45.05	63.43	0.65	0.73	0.60
Chart-SR (Ours)	**4B**	**0.71**	**0.62**	**0.80**	**68.31**	**47.68**	**66.15**	**0.69**	**0.80**	**0.62**
Improvement (%)		+26.79	+31.90	+21.20	+2.95	+5.85	+4.29	+6.15	+9.59	+3.33

4.2 Implementation Details

We use Qwen3-VL-4B-Instruct Bai et al. (2025a) as the primary backbone and evaluate model-scale generalization on Qwen3-VL-8B-Instruct Bai et al. (2025a). The RL stage samples G=8 candidate responses per input, and adopts learning rate LR_{RL}=5e-5, KL coefficient β=0.02, and we use LoRA fine-tuning Hu et al. (2021). Training is implemented with Transformers and TRL, with vLLM Kwon et al. (2023) used for RL sampling and inference. vLLM enables high-throughput generation through continuous batching and efficient memory management. We also incorporate a subset of single image data to improve training robustness. Experiments are conducted on NVIDIA A800-80GB GPU using mixed-precision training and gradient checkpointing.

4.3 Main Results

4.3.1 General Chart Understanding

Table 2 compares Chart-SR with three groups of models. The first group contains proprietary MLLMs, including GPT-4o, Gemini 2.0, and Claude 3.5 Sonnet. The second group contains general-purpose open-source models and chart models, including Qwen2-VL Wang et al. (2024),Qwen2.5-VL Bai et al. (2025b), QvQ-Preview-72B, InternVL2.5 Chen et al. (2024b), ChartGemma Masry et al. (2025), ChartSketcher Huang et al. (2026), and ChartPoint Xu et al. (2025b).

As shown in Table 2, Chart-SR achieves better performance on ChartQA, ChartBench and Evo-ChartQA, respectively. The gains are particularly pronounced on ChartBench, suggesting that the fine-grained visual perception and evidence-responsive reasoning learned from intervention-based training transfer to real-world charts, unannotated data points, and complex scientific figures.

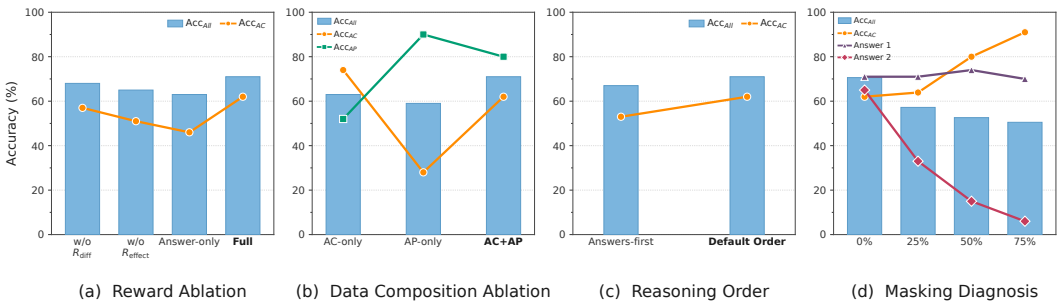


Figure 3: Ablation and diagnostic analyses of Chart-SR. (a) Removing R_{diff} or R_{effect} degrades performance, while the full reward design performs best. (b) AC-only and AP-only training exhibit complementary biases, whereas mixed AC+AP training achieves the best overall balance. (c) The default reasoning order outperforms Answers-first, especially on answer-changing cases. (d) Increasing masking on the perturbed chart (I_2) leaves Answer 1 largely stable but substantially degrades Answer 2 and overall accuracy, confirming reliance on the perturbed visual evidence.

4.3.2 Results on Chart-Perturb and Reasoning Benchmarks

Table 3 reports results on representative chart reasoning benchmarks, such as Chart-Perturb, MultiChart-QA, and RobustCQA. On Chart-Perturb, although Qwen3-VL-4B achieves reasonable effect prediction accuracy, its lower $\mathrm{Acc}AC$ and $\mathrm{Acc}AP$ indicate that correctly answering individual charts does not necessarily imply understanding how visual interventions affect the answers. In contrast, Chart-SR achieves relative gains of 26.79%, 31.90%, and 21.20% on $\mathrm{Acc}All$, $\mathrm{Acc}AC$, and Acc_{AP}, respectively. Chart-SR also maintains competitive performance on MultiChart-QA, with improvements on Sequential and Comparative, while achieving consistent gains of 6.15%, 9.59%, and 3.33% on the Overall, Augmented, and Human splits of RobustCQA. These results show that Chart-SR improves both intervention-sensitive reasoning and general robust chart reasoning.

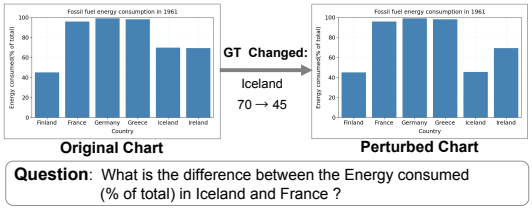
4.4 Ablation Study

Multi-Component Rewards. To evaluate the contribution of the difference-identification and effect rewards, we keep the 5-tag output format and all other training settings unchanged, while removing R_{diff}, R_{effect}, or both. The latter forms the Answer-only variant. As shown in Fig. 3(a), removing either reward consistently degrades both Acc_{All} and Acc_{AC}. Removing R_{effect} causes a larger performance drop than removing R_{diff}, indicating that explicitly supervising whether a visual change affects the answer is particularly important. Removing both rewards further reduces performance, with the Answer-only variant achieving the lowest results on both metrics. In contrast, the full reward design achieves the best Acc_{All} and Acc_{AC}, demonstrating the complementary benefits of the two reward components.

AC/AP Data Composition. To analyze the effect of training-data composition, we compare AC-only, AP-only, and mixed AC+AP training. As shown in Fig. 3(b), AC-only performs well on AC cases but remains inferior to mixed training overall, while AP-only better preserves answers but performs poorly when updates are required. In contrast, AC+AP training achieves the best overall performance, showing that reliable chart reasoning benefits from jointly learning when to update and when to preserve answers.

4.5 Analysis

Reasoning-Order Analysis. The default Chart-SR reasoning order follows *reasoning → original answer → difference identification → effect prediction → perturbed answer*. We compare it with an Answers-first variant, which generates both answers before identifying the difference and predicting its effect. As shown in Fig. 3(c), answers-first reduces Acc_{All} from 0.71 to 0.67, while Acc_{AC} drops more noticeably from 0.62 to 0.53. This larger degradation on AC samples suggests that identifying



Model	Ans1 (on original)	Difference (perceived)	Effect (on question-answer)	Ans2 (on perturbed)
Qwen3-VL-4B	25	iceland: 70→45	Preserved ✗	25 ✗
Chart-SR (Ours)	25	iceland: 70→45	Changed ✓	50 ✓

***Note**: (1) France is 95. (2) According to general chart practice, all answers have a 5% tolerance.*

Figure 4: Case study on PlotQA Methani et al. (2020) (`image_index:154585`). Chart-SR correctly identifies the question-relevant visual change and updates the answer accordingly, while the Qwen3-VL-4B backbone incorrectly preserves the original answer.

Table 4: Effect of SFT initialization before GRPO. The RL-only training performs better.

Training Strategy	Chart-Perturb (All)	Chart-Perturb (AC)	ChartQA	ChartBench
SFT + RL	0.63	0.55	0.82	62.70
RL-only	**0.71**	**0.62**	**0.86**	**68.60**

visual changes and judging their effects before generating the perturbed answer helps the model better determine when an answer should be updated.

Perturbed-Chart Masking Diagnosis. To verify whether the model genuinely reads the perturbed chart (I_2), we keep the original chart (I_1) and the question unchanged, while randomly masking 25%–75% of I_2 with white patches. As shown in Fig. 3(d), as the masking ratio increases, Answer 1, which depends only on I_1, remains largely stable, whereas Answer 2, which depends on I_2, degrades substantially, leading to a gradual decrease in Acc_{All}.

Interestingly, Acc_{AC} increases with heavier masking, as the model tends to interpret the masking itself as a visible chart change and thus becomes more likely to predict an answer change. Since all factors except I_2 remain unchanged, these systematic variations can only arise from the visual content of I_2. This confirms that both the second-chart answer and the intervention-effect judgment genuinely rely on reading perturbed chart (I_2), rather than on question templates or learned priors.

Effect of SFT Initialization. We further examine whether SFT initialization is necessary before GRPO. As shown in Table 4, direct GRPO performs better on both intervention and general benchmarks. This suggests that the pretrained backbone can acquire the required structured behavior directly from verifiable rewards. SFT on limited fixed trajectories may introduce additional imitation bias, while direct RL retains greater policy flexibility. This shows that Chart-SR can work well without SFT initialization.

4.6 Case Study

Figure 4 presents a representative case from PlotQA (`image_index:154585`), where question-relevant visual evidence is modified and the answer should be updated accordingly. Chart-SR correctly identifies the change, determines that it affects the answer to the question, and updates the prediction accordingly. In contrast, although Qwen3-VL-4B perceives the chart change, it incorrectly preserves the original answer, indicating weaker utilization of question-relevant evidence.

5 Conclusion

In this work, we propose Chart-SR, an evidence sensitivity reasoning framework for reliable chart understanding. By constructing chart pairs and optimizing MLLMs with GRPO and verifiable rewards, Chart-SR enables models to identify visual changes, judge their effects, and update or preserve answers accordingly. Experiments show that Chart-SR improves both general chart reasoning and perturbed consistency, providing an effective approach toward more reliable chart reasoning.

AI USE STATEMENT

In this work, we used generative AI tools to assist with paper writing and language polishing, including limited assistance with parts of the Introduction. AI assistance was primarily used to improve grammar, wording, clarity, and readability. Generative AI tools were not used to generate experimental results or to replace the authors' judgment regarding the research conclusions.

Notably, **no generative AI tools were used** to generate, annotate, or construct our Chart Sensitive Dataset (CS-Dataset) used for GRPO training. Instead, the CS-Dataset was entirely generated and computed through programmatic procedures based on chart annotations.

All AI-assisted text was manually reviewed, verified, and revised by the authors as necessary. The authors take full responsibility for the final content of this work.

REFERENCES

Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. Reading and reasoning over chart images for evidence-based automated fact-checking. In Andreas Vlachos and Isabelle Augenstein (eds.), *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 399–414, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.30. URL https://aclanthology.org/2023.findings-eacl.30/.

Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. Chartcheck: Explainable fact-checking over real-world chart images, 2024. URL https://arxiv.org/abs/2311.07453.

Elmira Amirloo, Jean-Philippe Fauconnier, Christoph Roesmann, Christian Kerl, Rinu Boney, Yusu Qian, Zirui Wang, Afshin Dehghan, Yinfei Yang, Zhe Gan, et al. Understanding alignment in multimodal llms: A comprehensive study. *arXiv preprint arXiv:2407.02477*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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. *arXiv preprint arXiv:2502.13923*, 2025b.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and Benyou Wang. Towards injecting medical visual knowledge into multimodal LLMs at scale. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7346–7370, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.418. URL https://aclanthology.org/2024.emnlp-main.418/.

Lei Chen, Xuanle Zhao, Zhixiong Zeng, Jing Huang, Liming Zheng, Yufeng Zhong, and Lin Ma. Breaking the sft plateau: Multimodal structured reinforcement learning for chart-to-code generation. *arXiv preprint arXiv:2508.13587*, 2025.

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, 2026. URL https://arxiv.org/abs/2507.15509.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Xinze Guan, and Xin Wang. Grit: Teaching mllms to think with images. *Advances in Neural Information Processing Systems*, 38:116522–116543, 2026.

Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. *arXiv preprint arXiv:2501.05452*, 2025.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Muye Huang, Han Lai, Xinyu Zhang, Wenjun Wu, Jie Ma, Lingling Zhang, and Jun Liu. Evochart: A benchmark and a self-training approach towards real-world chart understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3680–3688, 2025a.

Muye Huang, Lingling Zhang, Jie Ma, Han Lai, Fangzhi Xu, Yifei Li, Wenjun Wu, Yaqiang Wu, and Jun Liu. Chartsketcher: Reasoning with multimodal feedback and reflection for chart understanding. *Advances in Neural Information Processing Systems*, 38:70485–70515, 2026.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025b.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. Natural language understanding and inference with mllm in visual question answering: A survey. *ACM Comput. Surv.*, 57(8), March 2025. ISSN 0360-0300. doi: 10.1145/3711680. URL https://doi.org/10.1145/3711680.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Eunsoo Lee, Jeongwoo Lee, Minki Hong, Jangho Choi, and Jihie Kim. Visdot: Enhancing visual reasoning through human-like interpretation grounding and decomposition of thought. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 610–640, 2026.

Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. Deplot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10381–10399, 2023.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024b.

Ahmed Masry, Jia Qing Tan, Shafiq Joty, Enamul Hoque, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pp. 2263–2279, 2022.

Ahmed Masry, Mehrad Shahmohammadi, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. ChartInstruct: Instruction tuning for chart comprehension and reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10387–10409, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.619. URL https://aclanthology.org/2024.findings-acl.619/.

Ahmed Masry, Megh Thakkar, Aayush Bajaj, Aaryaman Kartha, Enamul Hoque, and Shafiq Joty. Chartgemma: Visual instruction-tuning for chart reasoning in the wild. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pp. 625–643, 2025.

Damien Masson, Sylvain Malacria, Daniel Vogel, Edward Lank, and Géry Casiez. Chartdetective: Easy and accurate interactive data extraction from complex vector charts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581113. URL https://doi.org/10.1145/3544548.3581113.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pp. 1527–1536, 2020.

Srija Mukhopadhyay, Adnan Qidwai, Aparna Garimella, Pritika Ramu, Vivek Gupta, and Dan Roth. Unraveling the truth: Do VLMs really understand charts? a deep dive into consistency and robustness. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16696–16717, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.973. URL https://aclanthology.org/2024.findings-emnlp.973/.

Hongkun Pan, Yuwei Wu, Wanyi Hong, Shenghui Hu, Qitong Yan, Yi Yang, Rufei Han, Changju Zhou, Minfeng Zhu, Dongming Han, et al. Chart-fr1: Visual focus-driven fine-grained reasoning on dense charts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26285–26294, 2026.

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

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.

Manan Suri, Puneet Mathur, Nedim Lipka, Franck Dernoncourt, Ryan A Rossi, and Dinesh Manocha. Chartlens: Fine-grained visual attribution in charts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22447–22462, 2025.

Alexander Vogel, Omar Moured, Yufan Chen, Jiaming Zhang, and Rainer Stiefelhagen. Refchartqa: Grounding visual answer on chart images through instruction tuning. In *International Conference on Document Analysis and Recognition*, pp. 523–537. Springer, 2025.

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

Renqiu Xia, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Botian Shi, Junchi Yan, and Bo Zhang. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *IEEE Transactions on Image Processing*, 2025.

Tianrun Xu, Haoda Jing, Ye Li, Yuquan Wei, Jun Feng, Guanyu Chen, Haichuan Gao, Tianren Zhang, and Feng Chen. Defacto: Counterfactual thinking with images for enforcing evidence-grounded and faithful reasoning. *arXiv preprint arXiv:2509.20912*, 2025a.

Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. Chartbench: A benchmark for complex visual reasoning in charts. *arXiv preprint arXiv:2312.15915*, 2023.

Zhengzhuo Xu, SiNan Du, Yiyan Qi, Siwen Lu, Chengjin Xu, Chun Yuan, and Jian Guo. Chartpoint: Guiding mllms with grounding reflection for chart reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 426–436, 2025b.

Zhangyuan Yu, Wanran Sun, Guangjing Yang, Xiaohu Wu, and Qicheng Lao. Cfpo: Counterfactual policy optimization for multimodal reasoning. *arXiv preprint arXiv:2606.23206*, 2026.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1859–1869, October 2025.

Xin Zhang, Xingyu Li, Rongguang Wang, Ruizhong Miao, Zheng Wang, Yuying Wang, Dan Roth, and Chenyang Li. Chart-RL: Generalized chart comprehension via reinforcement learning with verifiable rewards. In Vivek Gupta, Kaize Ding, Harsha Kokel, Yue Zhao, Amit Agarwal, Yu Wang, Michael Glass, Yu Zhang, Kavitha Srinivas, Xiusi Chen, Oktie Hassanzadeh, Qi Zhu, Shuaichen Chang, and Yuan Luo (eds.), *Proceedings of the First Workshop on Structured Understanding, Retrieval, and Generation in the LLM Era (SURGeLLM 2026)*, pp. 107–118, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-406-4. doi: 10.18653/v1/2026.surgellm-1.6. URL https://aclanthology.org/2026.surgellm-1.6/.

Kesen Zhao, Beier Zhu, Qianru Sun, and Hanwang Zhang. Unsupervised visual chain-of-thought reasoning via preference optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2303–2312, 2025.

Zifeng Zhu, Mengzhao Jia, Zhihan Zhang, Lang Li, and Meng Jiang. MultiChartQA: Benchmarking vision-language models on multi-chart problems. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11341–11359, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.566. URL https://aclanthology.org/2025.naacl-long.566/.

A Appendix

A.1 Details of Controlled Chart Interventions

Table 5 provides detailed definitions of the ten controlled chart interventions used in our data construction. Semantics-Changed (SC) interventions modify the underlying chart semantics, while their effects on the answer depend on whether the modified evidence is relevant to the question. Semantics-Preserved (SP) interventions only alter visual presentation and preserve the answer by construction. AC and AP denote Answer Changed and Answer Preserved, respectively.

Table 5: Detailed taxonomy of controlled chart interventions. SC denotes Semantics-Changed interventions, while SP denotes Semantics-Preserved interventions. The answer effect indicates whether each intervention produces Answer Changed (AC) or Answer Preserved (AP) samples in our dataset.

Type	Intervention	Description	Answer Effect
Semantics-Changed (SC)	SC-1 Data value	Modify the numerical value of a selected data point while keeping the remaining chart structure unchanged.	AC / AP
	SC-2 Data swap	Swap the values of two data points, changing their numerical associations while preserving other chart elements.	AC / AP
	SC-3 Global scaling	Multiply chart values by a common scaling factor, systematically changing their absolute magnitudes while preserving relative proportions.	AC / AP
	SC-4 Series removal	Remove one data series from the chart while leaving the remaining series and chart structure unchanged.	AC / AP
	SC-5 Legend label	Modify or swap legend labels, thereby changing the semantic mapping between series identities and their corresponding visual data.	AC / AP
Semantics-Preserved (SP)	SP-1 Series color	Change the color of one or more data series without modifying their values, labels, or semantic identities.	AP
	SP-2 Axis range	Adjust the displayed axis range while keeping all underlying data values and their semantic relationships unchanged.	AP
	SP-3 Category order	Reorder the displayed categories while preserving the original correspondence between each category and its values.	AP
	SP-4 Grid	Add, remove, or modify grid lines without changing any chart data, labels, or semantic mappings.	AP
	SP-5 Legend position	Move the legend to a different position while keeping its labels, colors, and series mappings unchanged.	AP