Diagnosing and Mitigating Structural Pathologies in Mathematical Reasoning of Large Language Models

Anonymous ACL submission

Abstract

Despite recent progress, large language models (LLMs) for mathematical reasoning often exhibit fragile behaviors, where correct answers are produced despite invalid or incoherent intermediate reasoning. We identify two recurring structural pathologies in Chain-of-Thought (CoT) reasoning: disconnected steps, where intermediate results are not reused, and weak logical flow, where steps are loosely or incorrectly linked yet still yield correct answers. These failures are difficult to address under outcome-only supervision. To mitigate these issues, we propose the Graph-structured Stepwise Reasoning Framework (GSRF), which reformulates implicit CoT into a Graph-structured Stepwise CoT (GS-CoT) that makes inter-step dependencies explicit. Building on this structure, we introduce Graph-guided Group Relative Policy Optimization (G-GRPO), incorporating process-level rewards that encourage step reuse and alignment with the final answer. Extensive experiments on both textual and multimodal mathematical reasoning benchmarks demonstrate that GSRF achieves competitive performance while producing more faithful, coherent, and structurally grounded reasoning traces.

1 Introduction

LLMs have achieved impressive progress on mathematical and geometric reasoning (Yan et al., 2025; Chen et al., 2025). Yet their reliability remains fragile, i.e., final-answer correctness is often decoupled from the validity of the underlying reasoning process. A trustworthy solution should satisfy a simple contract, i.e., each intermediate step should produce a result that is subsequently reused to support the final conclusion. In practice, standard CoT (Wei et al., 2022; Wang et al., 2022; Kojima et al., 2022) frequently violates this contract, i.e., the models may reach the correct answer while producing traces that are redundant, partially hallucinated, or logically disconnected, resulting in a convincing appearance of reasoning without verifiable structural support.

We argue that a major source of this unreliability is the implicit, free-form nature of standard CoT. Most current paradigms treat reasoning as a flat sequence of text, which obscures the dependency structure that makes a derivation checkable. As shown in Fig 1, this implicit CoT often exhibits two recurring structural pathologies: (i) *Disconnected Steps*, where intermediate results are introduced but never used (Turpin et al., 2023; Lanham et al., 2023); and (ii) *Weak Logical Flow*, where the intermediate steps exhibit unclear or even incorrect logical transitions, but lead to the correct answers (Lyu et al., 2023; Yee et al., 2024). Importantly, these pathologies are difficult to penalize under outcome-only supervision: as long as the final answer is correct, optimization signals provide little guidance on which steps are essential and which are merely spurious or redundant (Lightman et al., 2023).

To make reasoning verifiable and optimizable, we propose to capture reasoning faithfulness through explicit graph structure. We transform a solution into a Directed Dependency Graph (DDG), where nodes represent reasoning steps and edges encode variable-level reuse between steps. This view yields two efficiently graph-computable objectives that directly target the above failures: (i) *inter-step dependency*, encouraging steps to actually build on prior results; and (ii) *global step-to-answer support*, requiring that every step contributes (directly or indirectly) to the derivation of the final answer.

Based on these insights, we introduce a Graph-structured Stepwise Reasoning Framework (GSRF). GSRF replaces implicit CoT with an explicit and dependency-aware CoT called Graph-structured Stepwise CoT (GS-CoT), where step i is represented as a tuple (I_i, O_i) consisting of its inputs I_i (from problem conditions or previous

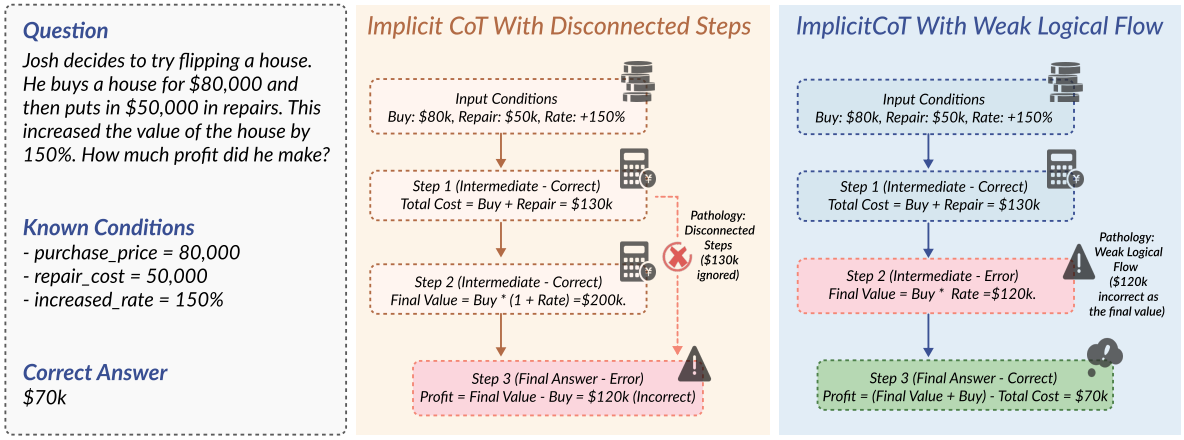

Figure 1: An example for solutions of implicit CoT with respectively disconnected steps and weak logical flow

outputs) and newly derived output O_i. To complement GS-CoT, we introduce Graph-guided Group Relative Policy Optimization (G-GRPO) to optimize the reasoning structure that moves beyond conventional outcome-level rewards (Zhang et al., 2025; Wang et al., 2025) and introduces two novel process-level reward signals: (i) *inter-step dependency reward* and (ii) *step-to-answer alignment reward* to assign process-level rewards for coherent reuse and global support. This directly discourages unused steps and promotes tight, reusable derivations. Experiments on both textual and multimodal benchmarks show that GSRF improves not only predictive accuracy but also the faithfulness and structural integrity of reasoning traces, suggesting a general principle for process optimization via explicit dependency structure. Our GSRF consists of two phases, and Fig. 2 shows an overview of our GSRF.

The **main contributions** of our work are threefold. First, we reveal two fundamental structural pathologies in implicit Chain-of-Thought for mathematical reasoning: (i) *disconnected steps* and (ii) *weak logical flow*. Second, we propose **GSRF** that replaces implicit CoT with **GS-CoT**, enabling the explicit construction of a **DDG** to model inter-step dependencies. Third, to further optimize reasoning structure, we introduce **G-GRPO** within GSRF, which deemphasizes outcome-level rewards and instead leverages two process-level rewards to guide dependency-aware and globally aligned reasoning.

2 Related Works

Mathematical Reasoning with Large Language Models Recent studies have shown that LLMs exhibit impressive yet fragile reasoning abilities in mathematical domains. Early works such as GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) benchmarked models' symbolic reasoning capacities, revealing their tendency to produce correct answers through flawed reasoning chains. To mitigate this, recent approaches such as DeepSeekMath-RL (Shao et al., 2024) and Math-Shepherd (Wang et al., 2024) leverage large-scale mathematical data and process-aware supervision to improve numerical precision and reasoning stability. However, these methods primarily focus on outcome-level or local step supervision, without explicitly modeling global reasoning dependencies, which limits interpretability and structural faithfulness.

Chain-of-Thought and Structured Reasoning Chain-of-Thought supervision has become a cornerstone for improving LLM reasoning by exposing intermediate steps. Subsequent works such as Tree-of-Thought (Yao et al., 2023) and Graph-of-Thought (Besta et al., 2024) attempted to model reasoning as structured search or graph traversal, encouraging exploration of diverse reasoning paths. Yet, most of these methods rely on heuristic sampling and lack an explicit optimization objective to align reasoning structure with answer correctness, often leading to redundant or inconsistent reasoning steps.

Reinforcement Learning for Reasoning Optimization. Reinforcement learning (RL) has been increasingly used to optimize reasoning behavior in LLMs. Methods like GRPO (Shao et al., 2024) and R1 (Guo et al., 2025) directly optimize reward signals derived from reasoning outcomes and format, effectively improving factual accuracy and reasoning stability. However, these trajectory-level or outcome-level objectives ignore inter-step depen-

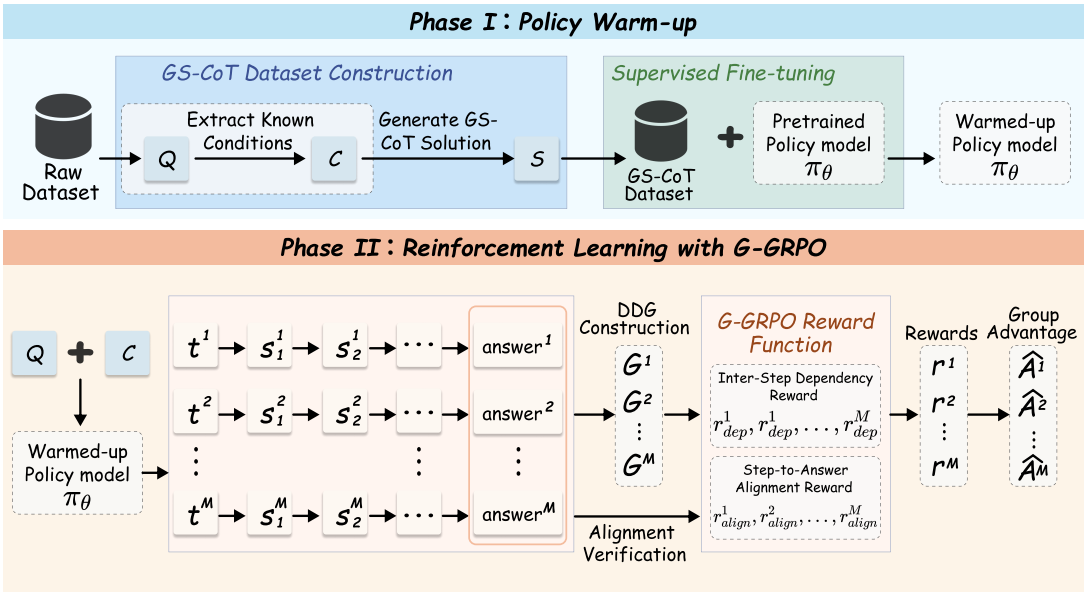

Figure 2: An overview of our GSRF consisting of two phases, i.e., Policy Warm-up and RL with G-GRPO.

dencies, which are particularly critical in mathematical reasoning where symbolic relations between steps dictate final correctness.

Multimodal and Structured Reasoning Extensions Recent efforts extend reasoning optimization into multimodal contexts, such as MathVista (Lu et al., 2023) and MM-ReAct (Yang et al., 2023). More recent work explores enhanced multimodal geometric reasoning, including unified vision–language pre-training methods for geometry such as GeoX (Xia et al., 2024), formal-verified multimodal problem generation via TrustGeoGen (Fu et al., 2025), and neuro-symbolic geometric data frameworks like NeSyGeo (Wu et al., 2025). Despite their progress, these approaches still ignore to explicitly model inter-step dependency and global structural consistency.

3 Methodology

3.1 Phase I: Policy Warm-up

GS-CoT Dataset Preparation. Before the policy warm-up phase, we construct a GS-CoT solution as the annotation for each mathematical problem. Specifically, given a mathematical problem Q, we first explicitly extract m known conditions, including numerical values, symbolic relations, and textual constraints, employing the DeepSeek-V3 model (Liu et al., 2024). These extracted conditions are denoted as $C = \{C_1, C_2, \ldots, C_m\}$. Next, given Q and its known conditions C, we employ DeepSeek-V3 again to generate a structured n-step solution $S = (s_1, s_2, \ldots, s_n)$, in which s_i $(i \leq n)$ strictly conforms to our predefined reasoning template with explicit inputs I_i and an output O_i where inputs I_i are subject to a following constraint:

$$I_i \subseteq \{C_1, C_2, \ldots, C_m\} \cup \{O_1, O_2, \ldots, O_{i-1}\}.$$

This constraint ensures that s_i depends only on the known conditions or the outputs of earlier steps, preserving dependency consistency in the reasoning process. Finally, we construct a high-quality GS-CoT dataset $\mathcal{D} = \{P_n | P_n = (Q_n, C_n, S_n)\}_{n=1}^{N}$ through a dual-review process combining the LLMs and human verification, which is then used for subsequent supervised fine-tuning. Note that, all GS-CoT annotations strictly satisfy the input–output dependency constraints by construction.

Supervised Fine-Tuning. After constructing the GS-CoT dataset, we train the policy model via supervised fine-tuning. This stage enables the model to learn explicit inter-step dependencies and to generate reasoning traces that strictly conform to the GS-CoT template. Formally, the supervised fine-tuning objective is defined as:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{T} \log \pi_\theta(y_t \mid y_{<t}, Q, C) \qquad (1)$$

where y_t denotes the t-th token in the reasoning sequence and Q and C are the input problem and known conditions, respectively. This stage establishes a strong initialization for subsequent policy refinement.

3.2 Phase II: RL with G-GRPO

To further refine the structural quality of the generated dependency graph and enhance the alignment between reasoning steps and final answers, we introduce G-GRPO, which augments the GRPO framework via two novel rewards: Inter-Step Dependency Reward and Step-to-Answer Alignment Reward rather than relying solely on the conventional outcome-based reward. This design encourages the model to focus on step-level structured reasoning rather than solely on the final answer. The Inter-Step Dependency Reward and the Step-to-Answer Alignment Reward are computed based on the Directed Dependency Graph (DDG). As illustrated in Fig. 3, we leverage the resulting graph structure to compute both rewards.

DDG Construction. Given a mathematical problem Q and its known conditions C , the warmed-up policy model generates its n-step GS-CoT solution $S = (s_1, s_2, \ldots, s_n)$. For the GS-CoT solution, we can define a DDG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = (v_0, v_1, \ldots, v_n)$. The initial node v_0 represents the set of known conditions C, and other nodes, e.g., $v_i (1 \leq i \leq n)$ represents the reasoning step s_i with its inputs I_i and output O_i. For an edge $e_{ij} \in \mathcal{E}$, $e_{ij} = 1$ if the input of v_j contains the output variable of v_i (i<j), that is $O_i \in I_j$; otherwise, $e_{ij} = 0$.

Inter-Step Dependency Reward. To encourage the model to produce reasoning chains that are both logically coherent and concise, we introduce a *dependency reward* r_{dep} based on the given DDG $\mathcal{G}$. Before calculating r_{dep}, we construct the adjacency matrix $A \in \{0,1\}^{N \times N} (N \geq 2)$ of DDG $\mathcal{G}$ with N nodes to describe inter-step dependencies and calculate the number of edges $|E(\mathcal{G})| = \sum_{i<j} A_{ij}$. We then calculate the inter-step dependency reward r_{dep}:

$$r_{\text{dep}} = \begin{cases} \lambda + 2(1-\lambda)P, & \text{if } |E(\mathcal{G})| \geq N-1, \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where λ is the normalization factor, usually taken as 0.6 and $P = \dfrac{|E(\mathcal{G})| - (N-1)}{(N-1)(N-2)}$. In addition, $r_{\text{dep}} = \lambda$ When $N = 2$.

This reward measures the proportion of valid dependency relations among all possible step pairs. A higher value indicates that reasoning steps are more tightly connected, reflecting a stronger structural dependency consistency.

Step-to-Answer Alignment Reward. While the dependency reward encourages dense and coherent inter-step connections, it does not explicitly guarantee that all reasoning steps contribute to the final answer. To explicitly enforce global step-to-answer consistency, we introduce a graph-theoretic *Step-to-Answer Alignment Reward* based on reachability in the reasoning graph.

Given a DDG $\mathcal{G}$ with $n+1$ reasoning nodes $(v_0, v_1, v_2, \ldots, v_n)$, where v_n corresponds to the final answer node, we define the reasoning steps are *aligned* with the final answer if and only if every preceding node v_i $(0 \leq i < n)$ can directly or indirectly influence v_n. Formally, this requires that for the node v_i, there exists at least one directed path from v_i to v_n in the graph $\mathcal{G}$. From a graph-theoretic perspective, this is equivalent to checking whether v_n is reachable from all other nodes in the graph $\mathcal{G}$. Let I_{reach} denote an indicator function defined as:

$$I_{\text{reach}} = \begin{cases} 1, & \text{if } \forall 0 \leq i < n,\ \exists \text{ a path } v_i \rightsquigarrow v_n, \\ 0, & \text{otherwise.} \end{cases}$$

Given the standard answer accuracy reward r_{acc}, the step-to-answer alignment reward is defined as:

$$r_{\text{align}} = \begin{cases} 1, & \text{if } I_{\text{reach}} = 1 \text{ and } r_{acc} = 1, \\ \gamma, & \text{if } I_{\text{reach}} = 1 \text{ and } r_{acc} \neq 1, \\ 0, & \text{if } I_{\text{reach}} = 0 \end{cases} \tag{3}$$

where $\gamma \in (0,1)$ is a constant, empirically taken as 0.4, which assigns partial credit to reasoning graphs whose steps are structurally aligned with the final answer but yield an incorrect prediction.

This reward enforces that the final answer must be supported by all preceding reasoning steps, either directly or indirectly, thereby discouraging redundant, disconnected, or logically irrelevant steps. By grounding the alignment criterion in reachability, the proposed reward provides a clear and interpretable mechanism for ensuring global coherence between intermediate reasoning and final prediction.

Policy Optimization. Given the inter-step dependency reward r_{dep} and the step-to-answer alignment reward r_{align} computed from the DDG $\mathcal{G}$, we define the overall reward for the i-th reasoning path as

$$r^i = r^i_{\text{dep}} + r^i_{\text{align}}.$$

By sampling M reasoning paths, we obtain a reward set $\{r^1, r^2, \ldots, r^M\}$, based on which we

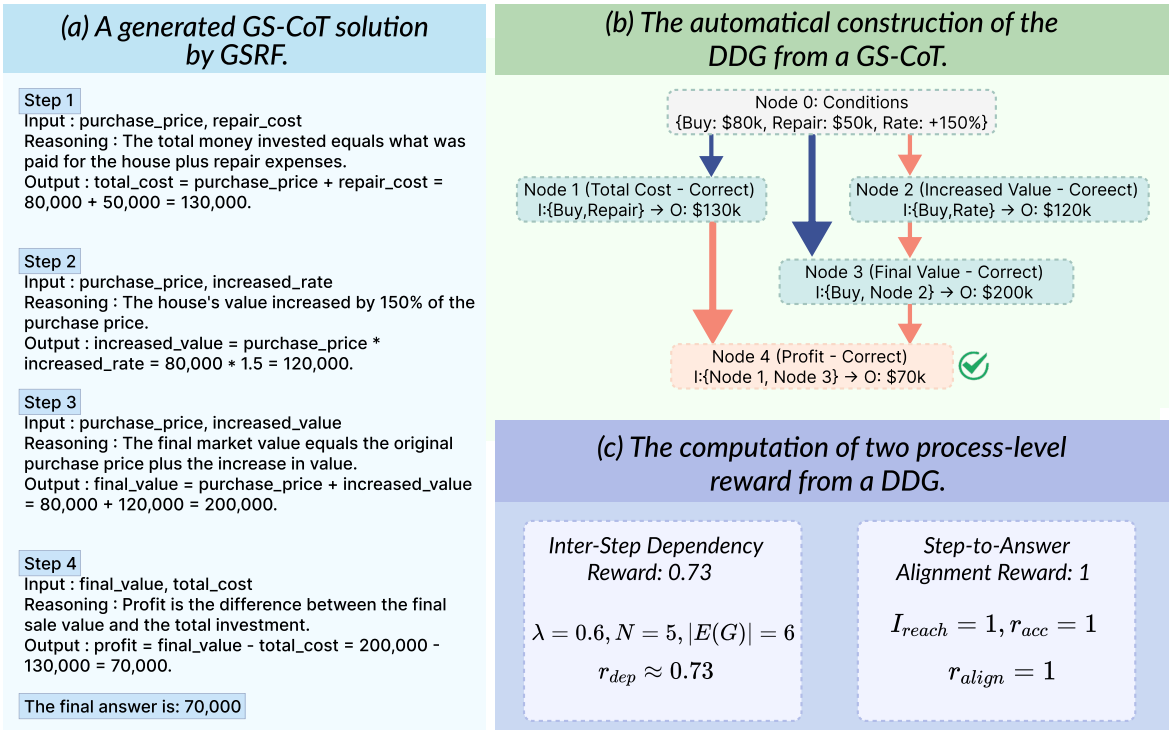

Figure 3: An overview on how to automatically construct a Directed Dependency Graph (DDG) for a GS-CoT data and calculate inter-step dependency reward and step-to-answer alignment reward (we use the same case in Fig. 1 and the two paths highlighted by orange arrows illustrate that every intermediate node in the DDG has a directed path to the final node.)

compute the relative advantage for each reasoning path as follows:

$$\hat{A}^i = \frac{r^i - mean(\{r^1, r^2, \ldots, r^M\})}{std(\{r^1, r^2, \ldots, r^M\})} \quad (4)$$

The mean group reward serves as a baseline, with $\hat{A}_i$ measures the deviation r_i deviates from the group average. We then optimize the policy model using the following loss function:

$$\mathcal{L}_{\text{G-GRPO}} = -\mathbb{E}_{Q \in D}\left[\frac{1}{M}\sum_{i=1}^{M}\left(\frac{\pi_\theta(c^i|Q)}{\pi_{\theta_{\text{old}}}(c^i|Q)}\right)\hat{A}^i - \beta D_{\text{KL}}\left(\pi_\theta \| \pi_{\text{ref}}\right)\right] \quad (5)$$

where a KL-divergence regularization term is applied to constrain the deviation of the policy model from a reference model. The reference model is initialized identically to the policy model and remains frozen throughout reinforcement learning. The KL divergence is estimated as follows:

$$D_{KL}(\pi_\theta \| \pi_{\text{ref}}) = \frac{\pi_{\text{ref}}(c^i|Q)}{\pi_\theta(c^i|Q)} - \log\frac{\pi_{\text{ref}}(c^i|Q)}{\pi_\theta(c^i|Q)} - 1 \quad (6)$$

The overall training pipeline is summarized in Appendix A.

4 Experiments

4.1 Experimental Setup

We conduct experiments on both textual and multimodal mathematical reasoning benchmarks to comprehensively evaluate the effectiveness of GSRF.

Textual Mathematical Reasoning. For pure textual reasoning, we evaluate our method on two widely used mathematical reasoning benchmarks, i.e., **GSM8K** and **MATH**, via Qwen2.5-3B and Qwen2.5-7B (Base Models) (Qwen et al., 2025). GSM8K focuses on grade-school level arithmetic, emphasizing numerical reasoning and stepwise calculation. In contrast, MATH contains competition-level problems spanning algebra, geometry, number theory, and calculus, requiring more complex symbolic manipulation and long-horizon reasoning. These two datasets together provide a comprehensive testbed for assessing structured reasoning ability under varying difficulty levels.

Multimodal Mathematical Reasoning. For multimodal reasoning, we adopt the GeoQA (Chen

et al., 2021) benchmark and the two baseline models, i.e., Qwen2.5-VL-3B and Qwen2.5-VL-7B (Bai et al., 2025). GeoQA consists of geometry problems that require joint reasoning over visual diagrams and textual descriptions, posing significant challenges for cross-modal alignment and multi-step deduction.

Evaluation Metrics. For all textual mathematical reasoning benchmarks, we adopt **exact-match accuracy** as the primary evaluation metric, following standard practice in prior work. A prediction is considered correct if and only if the final answer produced by the model exactly matches the ground-truth answer after normalization. Specifically, for GSM8K, we extract the final numerical result from the model output and compare it against the ground truth for MATH, correctness is determined based on symbolic equivalence, where answers are normalized to account for formatting variations such as whitespace, fractions, and equivalent algebraic expressions. For the multimodal benchmark, we evaluate GeoQA under two evaluation protocols—*open-ended* and *multiple-choice*—to better demonstrate the effectiveness of GSRF. In the open-ended setting, the model is provided with the visual diagram, the textual problem description, and the extracted known conditions, and is evaluated in the same manner as GSM8K by comparing the generated final answer against the ground truth. In the multiple-choice setting, the model is additionally given the full set of candidate answer options for each question. The predicted answer is matched against the provided choices, and the most consistent option is selected as the final output. Fig. 4 presents a quantitative comparison between Qwen2.5-VL-7B and GSRF-VL-7B under the multiple-choice setting on GeoQA.

4.2 Implementation Details

Implementation details can be found Appendix B.

4.3 Main Results and Analyses

Results on Textual Mathematical Reasoning. Table 1 reports the performance of our proposed GSRF on two representative textual mathematical reasoning benchmarks, GSM8K and MATH, compared with both closed-source and open-source models. GSRF consistently brings substantial improvements over the Qwen2.5 baselines on both benchmarks and across model scales. Notably, all GSRF results are obtained in a 0-shot setting, whereas most baseline models rely on few-shot prompting. This highlights the strong reasoning capability induced by our framework without requiring additional exemplars.

Table 1: Performance of GSRF on Qwen2.5-3B and Qwen2.5-7B on GSM8K and MATH datasets. Accuracy (%) across different models.

Model	GSM8K		MATH	
	Shot	Acc	Shot	Acc
Closed-Source Models				
GPT4	5	92.0	4	52.9
Gemini Ultra	-	94.4	4	53.2
Claude 3 Opus	0	95.0	0	60.1
Open-Source Models				
LLama3-70B-instruct	8	93.0	0	51.0
Qwen2.5-Math-72B	8	90.8	4	66.8
DeepSeekMath-RL-7B	-	88.2	-	51.7
Qwen2.5-3B (Baseline)	4	79.1	4	42.6
GSRF-3B (Ours)	0	**88.4**	0	**64.0**
Qwen2.5-7B (Baseline)	4	85.4	4	49.8
GSRF-7B (Ours)	0	**93.6**	0	**74.3**

On GSM8K, GSRF-3B achieves an accuracy of 88.4%, outperforming the Qwen2.5-3B baseline by +9.3%, despite the baseline using 4-shot prompting. Scaling up to 7B, GSRF-7B further improves the accuracy to 93.6%, surpassing its 4-shot baseline by +8.2%. On MATH, the gains are even more pronounced. GSRF-3B improves the accuracy from 42.6% to 64.0%, yielding a substantial +21.4% improvement over the baseline. Similarly, GSRF-7B boosts performance from 49.8% to 74.3% by +24.5%. These results not only surpass all compared open-source baselines, but also exceed several closed-source models, despite operating in a strictly 0-shot regime.

Results on Multimodal Mathematical Reasoning. Table 2 summarizes the results on the GeoQA benchmark under both multiple-choice and open-ended evaluation protocols. Under the multi-choice setting, GSRF brings significant gains over the Qwen2.5-VL baselines. Specifically, GSRF-VL-3B improves accuracy from 38.6% to 67.9% (+29.3%), while GSRF-VL-7B achieves 73.5%, outperforming its baseline by +14.2%. Notably, GSRF-VL-7B surpasses or matches strong open-source multimodal reasoning models such as MAVIS-7B (Zhang et al., 2024), G-LLaVA-13B (Gao et al., 2023), and Chimera-Reasoner-8B (Peng et al., 2025), despite operating in a strict 0-shot regime. Similar trends are observed in the more challenging open-ended setting. GSRF-VL-3B improves performance from 37.0% to 60.0%,

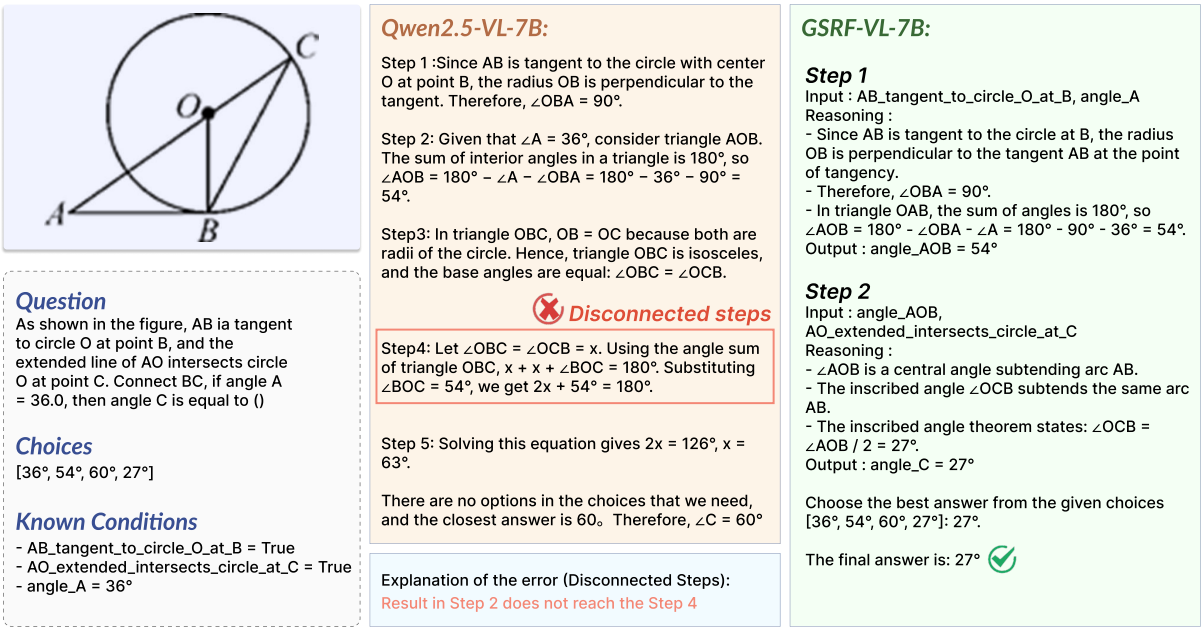

Figure 4: Quantitative comparison between Qwen2.5-VL-7B and GSRF-VL-7B (multi-choice) on GeoQA.

Table 2: Performance of GSRF on Qwen2.5-VL-3B and Qwen2.5-VL-7B on GeoQA dataset under multi-choice and open-ended evaluation protocols. Accuracy (%) across different models (0-shot).

Model	Accuracy
Multi-choice	
Closed-Source Models	
GPT-4o	58.9
Claude-3.5-Sonnet	65.1
Open-Source Models	
MAVIS-7B	68.3
G-LLaVA-13B	67.0
Chimera-Reasoner-8B	69.6
Qwen2.5-VL-3B (Baseline)	38.6
GSRF-VL-3B (Ours)	**67.9**
Qwen2.5-VL-7B (Baseline)	59.3
GSRF-VL-7B (Ours)	**73.5**
Open-ended	
Qwen2.5-VL-3B (Baseline)	37.0
GSRF-VL-3B (Ours)	**60.0**
Qwen2.5-VL-7B (Baseline)	46.2
GSRF-VL-7B (Ours)	**68.4**

and GSRF-VL-7B from 46.2% to 68.4%, demonstrating that the proposed framework effectively enhances structured geometric reasoning beyond answer selection and generalizes well to free-form solution generation.

4.4 Ablation Studies

To analyze the contribution of individual components in GSRF, we conduct ablation studies on reinforcement learning strategies, backbone models, reward design, and key hyperparameters. All ablation experiments follow the same training and evaluation protocols as the main experiments unless otherwise specified.

Comparison of RL Strategies. Table 3 compares different reinforcement learning strategies under the same backbone and training data. We first observe that policy warm-up consistently improves performance over models without warm-up on both GSM8K and MATH, indicating that graph-structured supervision provides a strong initialization for subsequent optimization. On GSM8K, our **G-GRPO** achieves the best performance across both model scales, demonstrating the benefit of explicitly incorporating graph-guided rewards for structured reasoning. On the more challenging MATH with the 3B model, G-GRPO (64.0%) slightly underperforms GRPO (65.9%) and GMPO (Zhao et al., 2025) (65.6%). We hypothesize that this effect is primarily due to the limited capacity of the smaller model, where enforcing stricter graph-level dependency constraints may overly constrain exploration and lead to suboptimal credit assignment in complex, long-horizon symbolic reasoning tasks. As model scale increases, this limitation is alleviated. For the 7B model, G-GRPO consistently outperforms both GRPO and GMPO on GSM8K and MATH, achieving 93.3% and 74.3% accuracy, respectively. Overall, these results suggest that graph-guided optimization is

Table 3: Comparison on **GSM8K** and **MATH**. Accuracy (%) across different reinforcement learning strategies

Model	Warm-up	GSM8K	MATH
Qwen2.5-3B	✗	79.1	42.6
Qwen2.5-3B	✓	80.6	57.4
Qwen2.5-3B+GRPO	✓	85.4	65.9
Qwen2.5-3B+GMPO	✓	84.7	65.6
Qwen2.5-3B+G-GRPO	✓	**88.4**	**64.0**
Qwen2.5-7B	✗	85.4	49.8
Qwen2.5-7B	✓	89.5	66.8
Qwen2.5-7B+GRPO	✓	90.0	71.4
Qwen2.5-7B+GMPO	✓	91.4	72.9
Qwen2.5-7B+G-GRPO	✓	**93.3**	**74.3**

Table 4: Performance on GSM8K with LLaMA3.1-8B as the backbone.

Model	Accuracy (%)
LLaMA3.1-8B-Instruct	84.5
GSRF-8B	**90.5**

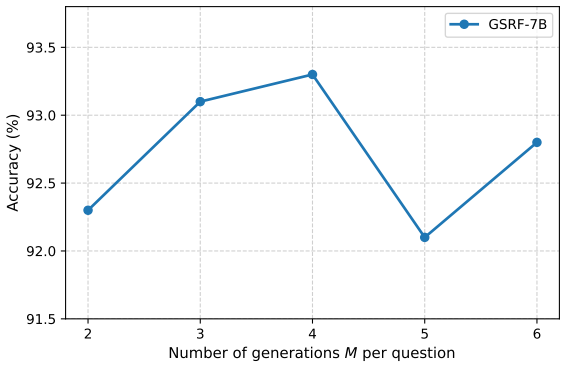

Figure 5: Parameter analysis of M. Accuracy (%) of GSRF-7B with Qwen2.5-7B as the backbone model under different M.

particularly effective when sufficient model capacity is available, allowing the policy to benefit from structured reasoning rewards without sacrificing optimization flexibility.

Effect of Model Backbone. Table 4 shows the performance of GSRF when applied to LLaMA3.1-8B as the backbone model on GSM8K. Compared to LLaMA3.1-8B-instruct, GSRF improves accuracy from 84.5% to 90.5%, yielding a gain of +6.0%. This result demonstrates that GSRF is not restricted to the Qwen family and can be effectively transferred to different backbone architectures, consistently enhancing mathematical reasoning performance.

Parameter Analyses. We conduct the parameter analysis on the number of generations M. As shown in Fig. 5, increasing M from 2 to 4 leads to consistent performance gains, while further increasing M yields marginal or no improvements. The results indicate that G-GRPO is not highly sensitive to the M, and a moderate value of M already provides competitive and stable performance.

5 Conclusion

We identify two structural pathologies in implicit Chain-of-Thought reasoning for mathematics: (i) *disconnected steps* and (ii) *weak logical flow*. To address these issues, we propose GSRF, a graph-structured reasoning framework that improves the faithfulness of Chain-of-Thought by making step dependencies explicit and optimizable. GSRF represents solutions as Graph-structured Stepwise CoT (GS-CoT) and constructs Directed Dependency Graphs (DDGs) to expose inter-step reuse and global support for the final answer. Building on this structure, we introduce Graph-guided Group Relative Policy Optimization (G-GRPO), which incorporates lightweight process-level rewards to penalize disconnected steps and weak logical flow beyond outcome-only supervision. Experiments on both textual and multimodal benchmarks demonstrate that GSRF achieves competitive performance while producing more faithful and structurally coherent reasoning traces.

Limitations

Our approach has several limitations. First, due to the structured nature of GS-CoT solutions, they cannot be reliably generated through prompt engineering alone. Instead, a GS-CoT dataset must be constructed in advance to warm up the model, which incurs additional annotation and training cost. Second, the two proposed rewards are complementary and need to be applied jointly; using either reward in isolation results in noticeable performance degradation. Third, the step-level reward design introduces the risk of reward hacking, where the model may exploit the reward signals without genuinely improving reasoning quality.

Further analysis and discussion are provided in Appendix C.

References

The claude 3 model family: Opus, sonnet, haiku.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and 1 others. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17682–17690.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. 2021. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523.

Zui Chen, Tianqiao Liu, Mi Tian, Qing Tong, Weiqi Luo, and Zitao Liu. 2025. Advancing mathematical reasoning in language models: The impact of problem-solving data, data synthesis methods, and training stages. *arXiv preprint arXiv:2501.14002*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daocheng Fu, Zijun Chen, Renqiu Xia, Qi Liu, Yuan Feng, Hongbin Zhou, Renrui Zhang, Shiyang Feng, Peng Gao, Junchi Yan, and 1 others. 2025. Trustgeogen: Scalable and formal-verified data engine for trustworthy multi-modal geometric problem solving. *arXiv preprint arXiv:2504.15780*.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and 1 others. 2023. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, and 1 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Tianshuo Peng, Mingsheng Li, Jiakang Yuan, Hongbin Zhou, Renqiu Xia, Renrui Zhang, Lei Bai, Song Mao, Bin Wang, Aojun Zhou, and 1 others. 2025. Chimera: Improving generalist model with domain-specific experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3011–3022.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan

Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439.

Teng Wang, Zhangyi Jiang, Zhenqi He, Shenyang Tong, Wenhan Yang, Yanan Zheng, Zeyu Li, Zifan He, Hailei Gong, Zewen Ye, and 1 others. 2025. Towards hierarchical multi-step reward models for enhanced reasoning in large language models. *arXiv preprint arXiv:2503.13551*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Weiming Wu, Jin Ye, Zi-kang Wang, Zhi Zhou, Yu-Feng Li, and Lan-Zhe Guo. 2025. Nesygeo: A neuro-symbolic framework for multimodal geometric reasoning data generation. *arXiv preprint arXiv:2505.17121*.

Renqiu Xia, Mingsheng Li, Hancheng Ye, Wenjie Wu, Hongbin Zhou, Jiakang Yuan, Tianshuo Peng, Xinyu Cai, Xiangchao Yan, Bin Wang, and 1 others. 2024. Geox: Geometric problem solving through unified formalized vision-language pre-training. *arXiv preprint arXiv:2412.11863*.

Yibo Yan, Jiamin Su, Jianxiang He, Fangteng Fu, Xu Zheng, Yuanhuiyi Lyu, Kun Wang, Shen Wang, Qingsong Wen, and Xuming Hu. 2025. A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11798–11827.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, and 1 others. 2024. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Evelyn Yee, Alice Li, Chenyu Tang, Yeon Ho Jung, Ramamohan Paturi, and Leon Bergen. 2024. Dissociation of faithful and unfaithful reasoning in llms. *arXiv preprint arXiv:2405.15092*.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, and 1 others. 2024. Mavis: Mathematical visual instruction tuning with an automatic data engine. *arXiv preprint arXiv:2407.08739*.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, Fang Wan, and Furu Wei. 2025. Geometric-mean policy optimization. *Preprint*, arXiv:2507.20673.

Algorithm 1 GSR Training Pipeline

Input: Pretrained policy model π_θ; GS-CoT dataset $\mathcal{D} = \{P_n\}_{n=1}^{N}$.
Output: Trained policy model π_θ

Policy Warm-up:
for iter $= 1$ to I **do**
  Sample P from $\mathcal{D}$
  Update policy model π_θ by Eq. (1)
end for
RL with G-GRPO:
for iter $= 1$ to I **do**
  Sample problem P from $\mathcal{D}$
  Generate M reasoning trajectories $\{t^i\}_{i=1}^{M}$ and Parse steps and construct dependency graph $\{\mathcal{G}^i\}_{i=1}^{M}$
  Compute inter-step dependency reward $\{r^i_{\text{dep}}\}_{i=1}^{M}$ by Eq. (2) and step-to-answer alignment reward $\{r^i_{\text{align}}\}_{i=1}^{M}$ by Eq. (3)
  Compute the relative advantage $\hat{A}^i$ by Eq. (4)
  Optimize Policy model π_θ by Eq. (5)–(6)
end for
  return policy model π_θ

A GSRF Training Pipeline

The overall training pipeline of GSRF is summarized in Algorithm 1.

B Implementation Details

For all datasets, we first employ **DeepSeek V3** to automatically extract the *known conditions* for each mathematical problem, including numerical constants, symbolic relations, and textual constraints. Based on these extracted conditions, DeepSeek V3 is further prompted to generate structured solutions that strictly follow our graph-structured reasoning template. This procedure is applied uniformly to all training data in both text-only and multimodal settings.

In policy warm-up phase, we fine-tune all backbone models using supervised learning on the generated graph-structured reasoning data. To ensure parameter-efficient adaptation across different model scales, we adopt **LoRA** for fine-tuning. Specifically, we apply LoRA to all eligible modules with rank 8 and scaling factor 16. The per-device training batch size is set to 8, and the learning rate is fixed at 5×10^{-5}.

In RL with G-GRPO phase, both the policy model and the reference model are initialized from the warmed-up checkpoint, but the reference model remains frozen throughout this phase. we further optimize the policy model using the proposed Graph-guided Group Relative Policy Optimization (G-GRPO). For each question, we sample 4 reasoning trajectories to form a comparison group, with the maximum response length set to 1024 tokens. The policy model is optimized with a learning rate of 1×10^{-6}, and the per-device training batch size is 32. The KL regularization coefficient between the policy and reference models is set to 0.001 to stabilize training while allowing sufficient exploration. All experiments are conducted on 8 RTX L40 48GB GPUs.

Fig. 4 shows the quantitative comparison between Qwen2.5-VL-7B and GSRF-VL-7B (multi-choice) on GeoQA.

C Additional Analysis and Discussion

This section provides a more detailed analysis of GSRF in relation to prior work discussed in the introduction, and further examines its limitations in a broader context.

Comparison with Prior Work. Recent advances in mathematical reasoning with large language models have largely focused on improving answer accuracy through larger-scale supervision, domain-specific fine-tuning, and reinforcement learning with outcome-level or locally defined process rewards. These approaches have demonstrated that optimizing reasoning trajectories can improve robustness and reduce superficial errors. However, most of them implicitly represent reasoning as a linear text sequence and do not explicitly model how intermediate steps depend on each other or collectively contribute to the final answer.

Several structured reasoning paradigms, such as tree-based or graph-based reasoning, attempt to address this issue by introducing explicit branching or search over reasoning paths. While effective in certain settings, these methods typically rely on heuristic exploration or external controllers, which increases inference-time complexity and makes optimization less direct. In contrast, our approach focuses on explicit structural supervision without introducing search, by replacing implicit CoT with GS-CoT and constructing a Directed Dependency Graph (DDG) over reasoning steps. This representation enables us to define step-level rewards that target global structural properties, such as dependency consistency and step-to-answer support, which are difficult to capture with token-level or outcome-only objectives.

At the same time, GSRF does not aim to fully verify the semantic correctness of each reasoning step. Instead, it emphasizes structural faithfulness, encouraging intermediate steps to be meaningfully

connected and functionally relevant to the final answer. From this perspective, GSRF should be viewed as complementary to existing methods that focus on correctness, efficiency, or exploration, rather than as a replacement for them.

Limitations. Despite its effectiveness, GSRF has several limitations. First, the framework relies on GS-CoT annotations to initialize training. Unlike standard CoT, GS-CoT cannot be reliably induced through prompting alone, and constructing such data—especially with human verification—introduces additional cost. This limits the immediate applicability of GSRF to domains where structured reasoning annotations are available or affordable. Second, the two step-level rewards used in G-GRPO capture different but complementary aspects of reasoning structure. The inter-step dependency reward encourages coherent dependency chains, while the step-to-answer alignment reward promotes global relevance. Empirically, removing either reward leads to performance degradation, indicating that each alone is insufficient. This coupling makes reward design less modular and may complicate extension to more complex or heterogeneous reasoning tasks. Third, as with other reinforcement learning approaches, GSRF is susceptible to reward exploitation. Because the rewards are defined over graph structure rather than semantic validity, models may learn to satisfy structural criteria in superficial ways. Although this risk is mitigated through joint supervision and regularization, it remains an inherent challenge of process-level optimization. Finally, enforcing explicit structural constraints may restrict exploration, particularly for smaller models or tasks requiring long and flexible reasoning chains. Our results suggest that this effect diminishes with increased model capacity, but it highlights a trade-off between structural guidance and exploratory freedom.

Overall, GSRF represents a step toward more structurally faithful and interpretable reasoning, while inheriting some limitations common to structured supervision and reinforcement learning. We view it as a complementary component that can be combined with semantic verification or external reasoning tools in future work.