Chart Deep Research in LVLMs via Parallel Relative Policy Optimization

Jiajin Tang[1,2], Gaoyang[1], Wenjie Wang[2], Sibei Yang[3], ✉, Xing Chen[1]

[1]ByteDance, [2]Shanghaitech University, [3]School of Computer Science and Engineering, Sun Yat-sen University

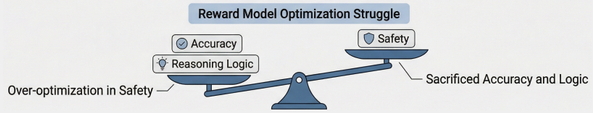
ICLR 2026

1. INTRODUCTION & MOTIVATION

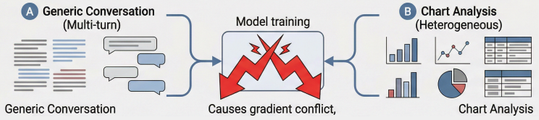
Chart Deep Research in Large Vision-Language Models (LVLMs) is the task of generating **in-depth, logical reports** and **forecasts** from **heterogeneous charts**, moving beyond simple chart question answering.

Current LVLMs **struggle** to provide logical, professional-grade chart analysis.

Multidimensional Reward Signals. "Deep research requires balancing accuracy, reasoning logic, and safety. A single reward model struggles to optimize these simultaneously, often leading to over-optimization in one dimension (e.g., safety) at the expense of others (e.g., accuracy)."

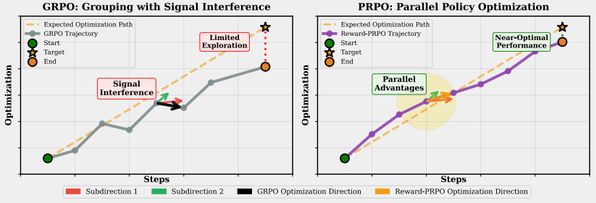


Heterogeneous Data Conflict. "Combining generic multi-turn data with diverse chart-specific data causes gradient conflict, hampering generic conversational ability."



2. PROBLEM FORMULATION & BASELINE

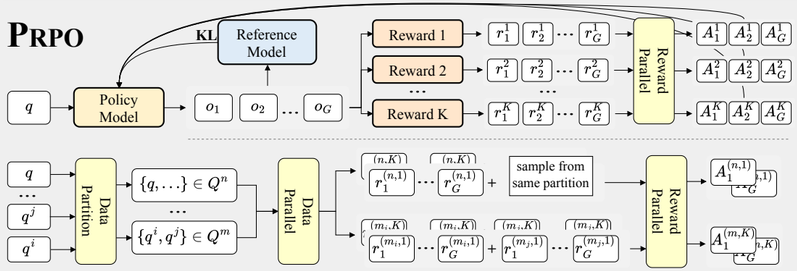
GRPO Loss: $$J_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\text{old}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} L_{\text{clip}}(r_{i,t}(\theta), \hat{A}_i) \right]$$



Reward scalarization compresses variability, weakens signals, and diminishes advantage discriminative power.

Simple-data dominate training, suppressing informative complex-data complex reasoning signals.

PRPO: PARALLEL RELATIVE POLICY OPTIMIZATION



3. PARALLEL RELATIVE POLICY OPTIMIZATION WITH REWARD & DATA PARALLELISM

We propose **PRPO**, a unified framework to address these bottlenecks using RLHF. The key idea is to parallelize and decouple the training process at **reward** and **data** level.

Reward-PRPO (Reward Parallelism)

Reward-PRPO manages multiple reward signals (e.g., Accuracy Reward, Reason Reward, Safety Reward). It optimizes them in parallel using independent relative policy updates for each dimension, avoiding the trade-off common in multi-objective optimization.

$$\hat{A}_i^{(k)} = \frac{R_i^{(k)} - \bar{R}^{(k)}}{\sigma^{(k)}}, \quad k = 1, 2, \ldots, K$$

$$J_{\text{Reward-PRPO}}(\theta) = \sum_{k=1}^{K}\lambda_k \mathbb{E}_{q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\text{old}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} L_{\text{clip}}(r_{i,t}(\theta), \hat{A}_i^{(k)}) \right]$$

Data-PRPO (Data Parallelism)

Data-PRPO enhances GRPO by implementing Capability-based Level Identifiers to partition training data into homogeneous reward distributions. By normalizing advantages within these partitions and implementing an Iterative Validation mechanism to relegate statistical outliers, Data-PRPO mitigates optimization conflicts arising from heterogeneous data types while maintaining global policy stability.

$$\hat{A}_i^{(m)} = \frac{R_i - \bar{R}^{(m)}}{\sigma^{(m)}},$$

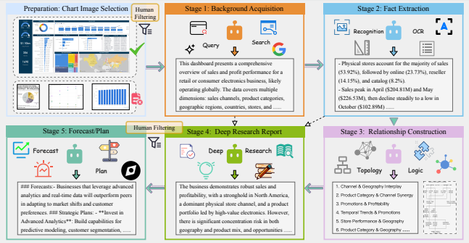
where m denotes the macro-partitioning of samples based on their underlying cognitive capabilities.

$$J_{\text{Data-PRPO}}(\theta) = \sum_{m=1}^{M_{\text{final}}}\lambda_m \mathbb{E}_{q\sim P(Q^{(m)}),\{o_i\}_{i\in\mathcal{G}_m}\sim\pi_{\text{old}}(\cdot|q)} \left[\frac{1}{|\mathcal{G}_m|}\sum_{i\in\mathcal{G}_m}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} L_{\text{clip}}(r_{i,t}(\theta), \hat{A}_i^{(m)}) \right]$$

Overall PRPO

$$J_{\text{PRPO}}(\theta) = \sum_{m=1}^{M_{\text{final}}}\lambda_m \sum_{k=1}^{K}\lambda_k \mathbb{E}_{q\sim P(Q^{(m)}),\{o_i\}_{i\in\mathcal{G}_m}\sim\pi_{\text{old}}(\cdot|q)} \left[\frac{1}{|\mathcal{G}_m|}\sum_{i\in\mathcal{G}_m}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} L_{\text{clip}}(r_{i,t}(\theta), \hat{A}_i^{(k,m)}) \right]$$

4. MCDR-BENCH: MULTIMODAL CHART DEEP RESEARCH BENCHMARK



We introduce MCDR-Bench, based on the Error Unique Principle. Evaluate deep research across five logical stages:
(1) **Background Acquisition**: retrieving domain-specific knowledge; (2) **Fact Extraction**: extracting atomic data elements; (3) **Relationship Construction**: modeling topological and logical connections; (4) **Deep Research Report Generation**: synthesizing comprehensive reports; (5) **Forecast/Plan**: proposing strategic recommendations. Human filtering ensures quality control throughout the process.

5. KEY RESULTS: TRAINING COMPARISON & FINAL PERFORMANCE

Main Result Table 1: Performance comparison on MCDR-Bench across different splits.

Model	BG	FE	RL	DR	F/P	Overall	Mean
GRPO	41.24	51.69	75.38	66.14	77.4	61.71	62.26
PRPO (Ours)	50.65	61.38	81.78	72.83	84.01	69.62	69.9
GRPO Think	43.13	48.77	77.75	70.28	81.02	63	63.99
PRPO Think (Ours)	**62.9**	**65.23**	**88.87**	**80.91**	**87.21**	**76.26**	**76.89**

Main Result Table 2: Performance comparison on ChartQAPRO across different splits.

Model	Factoid	MCQ	Conv.	FactChk	Hypo.	Overall	Mean
GPT-4o	35.76	46.72	34.75	45.49	28.91	37.67	38.33
Gemini-Flash-2.0	43.43	60.28	40.25	67.62	24.47	46.85	47.15
Qwen2.5-VL-7B	27.49	37.85	55.22	46.72	44.4	36.31	41.33
w/ GRPO	-	-	-	-	-	39.97	-
w/ PRPO (Ours)	36.24	50.47	49.63	53.28	53.69	42.95	47.69

Main Result Figure 1: Accuracy reward trajectories during training for different optimization strategies.

PRPO's robust generalization across diverse tasks validates our parallel optimization strategy.

6. CONCLUSION & SUMMARY

PRPO is the first unified framework to address the bottlenecks of deep chart research in LVLMs. Its parallel, decoupled RLHF architecture resolves reward/data conflicts and enables balanced multidimensional optimization. PRPO-trained models achieve expert-level, logical chart analysis, setting a new SOTA and narrowing the gap with commercial models.

- ✅ We systematically analyze the bottlenecks constraining chart deep research capability development.
- ✅ We propose PRPO, enabling coordinated development of complex analytical capabilities.
- ✅ We present MCDR-Bench, providing systematic evaluation of chart deep research capabilities.
- ✅ Together, our unified framework establishes a systematic pathway for advancing chart deep research capabilities.

Project Page