Diversity-Incentivized Exploration for Versatile Reasoning

Zican Hu[12*], Shilin Zhang[12*], Yafu Li[23†], Jianhao Yan[24], Xuyang Hu[2],
Leyang Cui[4], Xiaoye Qu[2], Chunlin Chen[1], Yu Cheng[3], Zhi Wang[12]

[1]Nanjing University [2]Shanghai AI Laboratory [3]The Chinese University of Hong Kong [4]Westlake University

Outline

- **Background:** Challenges in RL verify reasoning?
- **Motivation:** Why global diversity?
- **Method:** Diversity metrics, intrinsic reward, mitigation strategies
- **Research Question 1:** Does the diversity enhance reasoning capacity?
- **Research Question 2:** Does DIVER achieve broader and more effective exploration?
- **Research Question 3:** How to mitigate reward hacking caused by diversity?

1. Background

Entropy Collapse

Policy distribution becomes overly concentrated, reducing exploration

Insufficient Exploration

struggle to discover novel solution pathways

Reward Sparsity

Massive "reward deserts" in complex reasoning tasks

Local Optima

Current approaches focus on local diversity, missing global patterns

2. Motivation

Key Insight: Diversity Matters

Strong positive correlation between global diversity and reasoning capacity

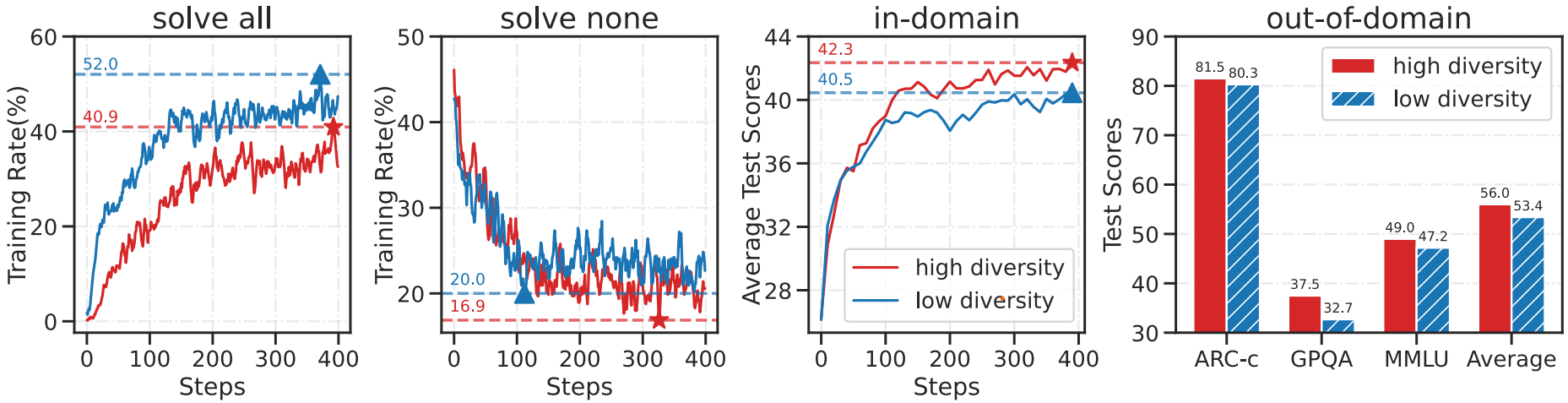
In-domain

High Diversity **42.3** → Low Diversity **40.5** **+1.8**

Out-of-domain

High Diversity **56.0** → Low Diversity **53.4** **+2.6**

High-diversity training achieves better generalization, especially on out-of-domain tasks



2. Motivation

Versatile Reasoning

Explore diverse solution patterns

Significantly improve deep exploration capabilities

Global vs. Local Exploration

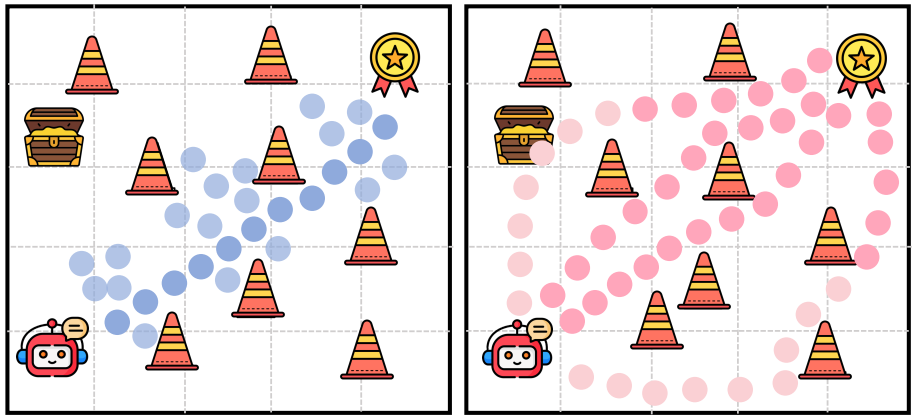
Sequence-level vs. Token-level

Deep exploration requires far-sighted, temporally-extended diversity

Intrinsic Motivation

Global diversity as intrinsic reward

Incentivize exploration in semantically structured space



(a) Local exploration (b) Global exploration (Ours)

3. Method

DIVER Framework

1 Diversity Metrics

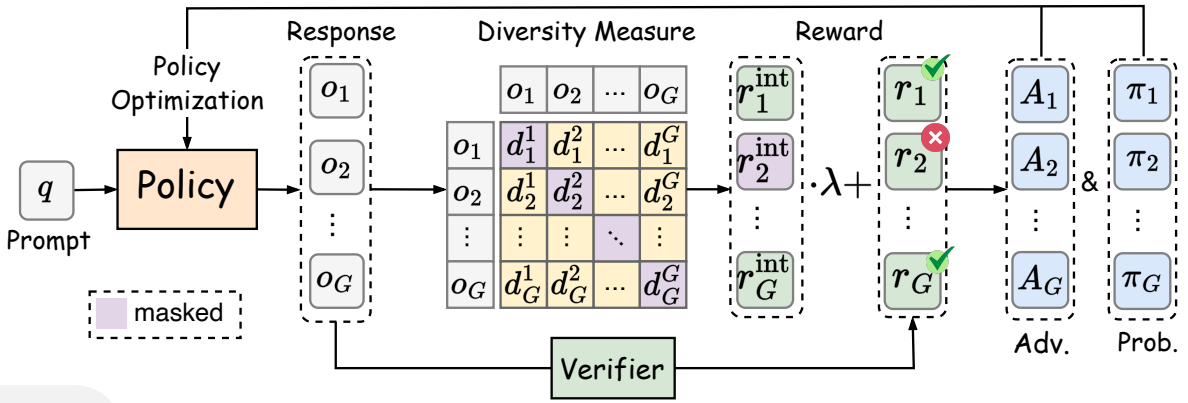
- Textual Diversity (TD): BLEU-based dissimilarity
- Equational Diversity (ED): Formula pattern uniqueness

$$\mathrm{TD}(o_i) = \frac{1}{G-1}\sum_{j\in[G]\setminus\{i\}}\Big(1-\mathrm{BLEU}(o_i,o_j)\Big).$$

$$\mathrm{ED}(o_i) = \frac{\big|\mathcal{F}(o_i)\setminus\mathcal{F}_{-i}\big|}{\big|\mathcal{F}(o_i)\big|}, \quad \text{if } |\mathcal{F}(o_i)|>0; \quad \text{or } 0, \text{ otherwise.}$$

2 Intrinsic Reward

- Augmented reward: $R' = R + \lambda \cdot R_{int}$
- Optimal policy invariance (Theorem 1)



3 Mitigating Reward Hacking

- Balanced Shaping: clip upper bound + decay λ
- Conditional Shaping: reward diversity only for correct answers

$$r_i^{\text{int}} = \text{clip}(r_i^{\text{int}}; 0, \sigma)$$

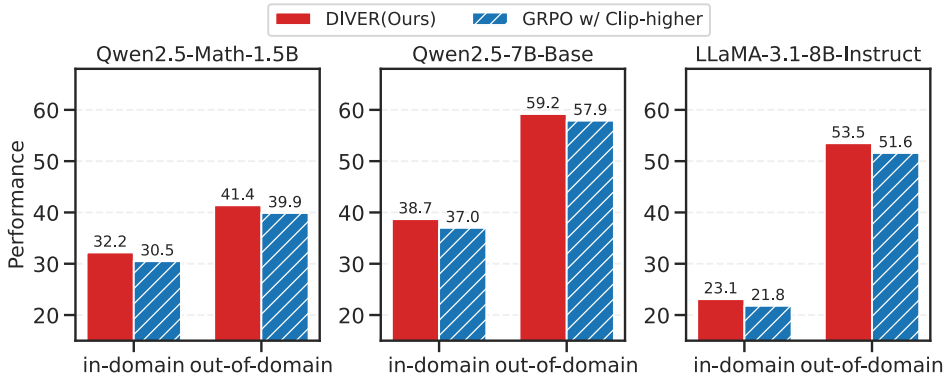
$$r_i' = r_i + \lambda \cdot r_{\text{int}}^i \cdot I(r_i)$$

4. Research Question

Can DIVER improve performance while maintaining effective global exploration and reliably extending to other models?

Yes! Consistent improvements across models and domains

Model	In-Domain Performance						Out-of-Domain Performance			
	AIME 24/25	AMC	MATH-500	Minerva	Olympiad	Avg.	ARC-c	GPQA*	MMLU-Pro	Avg.
Qwen2.5-Math-7B	11.8/6.3	43.1	56.8	16.9	25.4	26.7	38.1	12.2	31.5	27.3
Previous RLVR methods										
SimpleRL-Zoo	**25.2**/12.0	57.6	76.2	27.2	41.0	39.9	22.0	20.4	32.5	25.0
OpenReasoner-Zero	16.5/15.0	52.1	82.4	33.1	**47.1**	41.0	66.2	29.8	**58.7**	51.6
PRIME-Zero	17.0/12.8	54.0	81.4	**39.0**	40.3	40.7	73.3	18.2	32.7	41.4
Exploration RL Methods										
GRPO w/ Clip-higher	18.9/16.4	57.3	81.2	28.7	41.5	40.7	82.1	36.2	47.2	55.2
Entropy-RL	23.6/12.8	58.4	82.8	31.6	41.5	41.8	80.7	38.8	48.4	56.0
Pass@k Training	19.8/14.3	54.7	80.2	29.0	41.5	39.9	82.1	**44.4**	47.8	58.1
Our Methods										
DIVER-TD	22.5/**16.9**	59.4	82.2	27.9	44.7	42.3	**83.4**	42.3	49.5	**58.4**
DIVER-ED	20.9/15.7	**59.7**	**84.0**	31.6	46.1	**43.0**	**83.4**	36.2	49.9	56.5

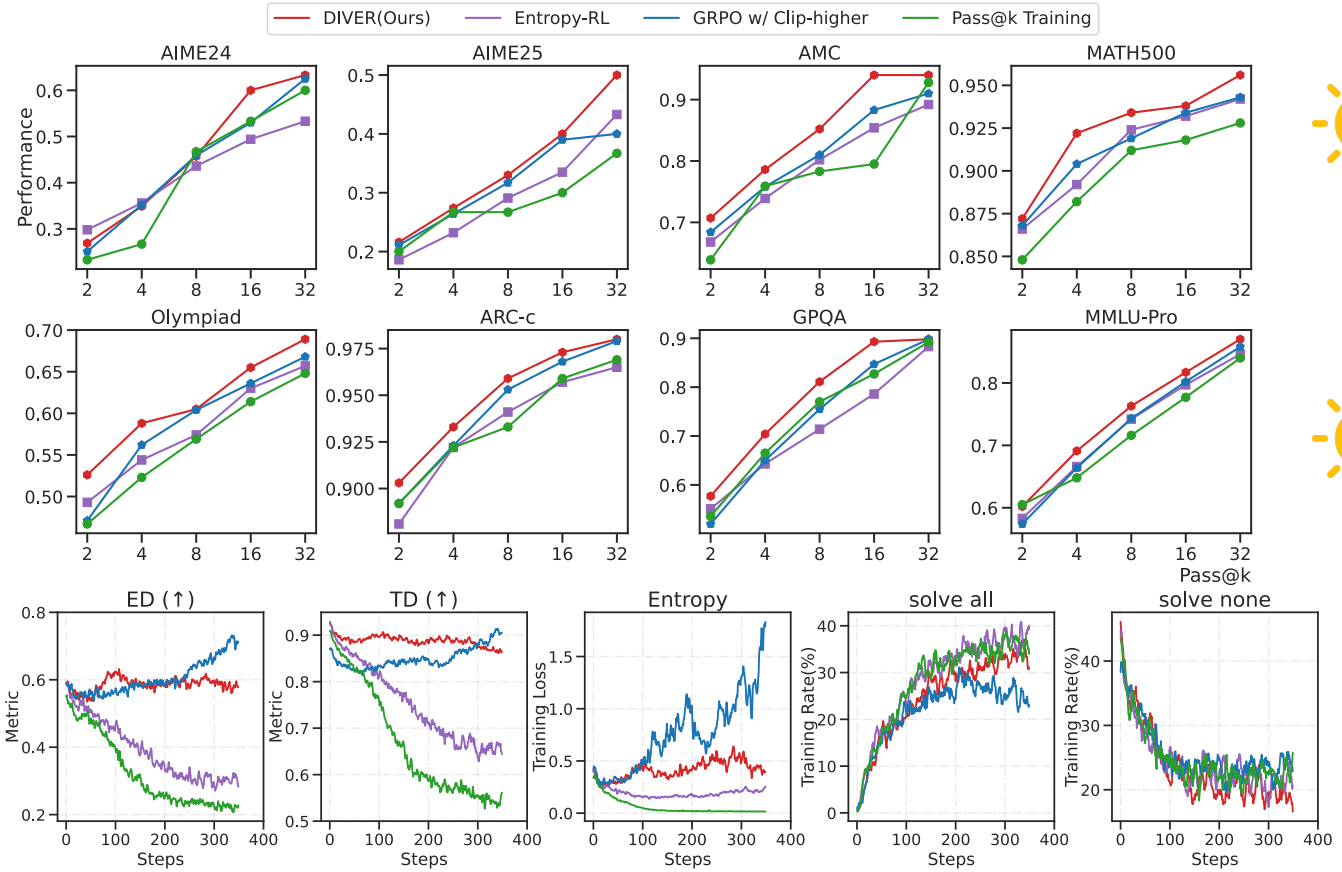


DIVER maintains effectiveness across in-domain and out-of-domain based on various model backbones

4. Research Question

Can DIVER achieve an effective and broader exploration scope that unlocks enhanced reasoning capacity?

Yes! Superior exploration with better Pass@k



DIVER unlocks higher reasoning scope with broader exploration

DIVER achieves optimal balance: high diversity + stable entropy + controlled exploration → broader reasoning scope

4. Research Question

Key Results Summary

Math Reasoning (In-Domain)

+1.2 points vs. exploration method (Avg. 43.0)

AIME24: 22.5 | AIME25: 16.9 | AMC: 59.7 | MATH-500: 84.0 | OlympiadBench: 46.1

Generalization (Out-of-Domain)

+6.8 points vs. non-exploration method (Avg. 88.4)

ARC-c: 83.4 (+17.2) | GPQA*: 42.3 (+13.3) | MMLU-Pro: 49.9 (+17.2)

Reasoning Capacity Scope

Superior Pass@k across all benchmarks

Consistent improvements from k=2 to k=32

4. Research Question

What is the appropriate configuration of DIVER?

Different design choices validated through ablation studies

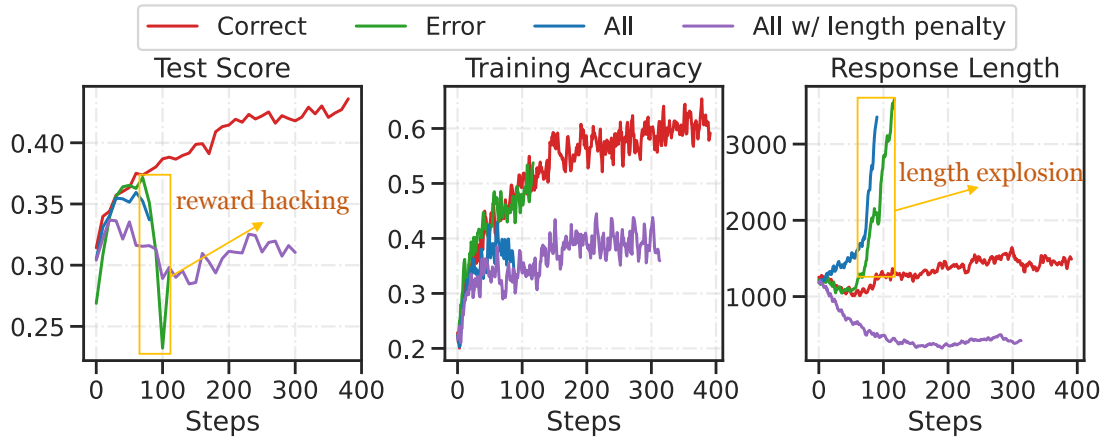
All (unconditional)
Diversity rewards for all responses
Reward hacking: exploits diversity and length explosion

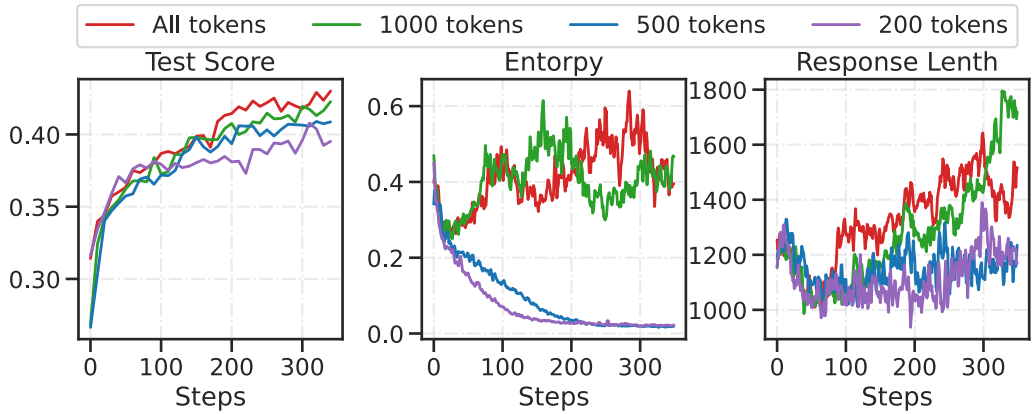
Error only
Diversity rewards only for error responses
Reward hacking: exploits diversity and length explosion

ALL (w/ length penalty)
Diversity rewards for all responses with length control
Reward hacking: exploits diversity and length explosion

Conditional Shaping (Correct only)
Diversity rewards for correct responses
Best test score (0.43) + Controlled length (1500)

Conditional & Balance & Global Diversity Reward Shaping





5. Summary

- Global sequence-level diversity is pivotal for versatile reasoning
- DIVER incentivizes deep exploration via diversity-based intrinsic rewards
- We proves Optimal policy invariance
- DIVER achieves effectiveness across in-domain and out-of-domain based on various model backbones
- DIVER Maintains high diversity with stable entropy, optimal exploration balance
- DIVER unlocks higher reasoning scope with broader exploration

🌟 DIVER is an efficient global exploration method in LLM reasoning

Paper: https://huggingface.co/papers/2509.26209 Code: https://github.com/NJU-RL/DIVER

Thank you!

Zican Hu[12*], Shilin Zhang[12*], Yafu Li[23†], Jianhao Yan[24], Xuyang Hu[2],
Leyang Cui[4], Xiaoye Qu[2], Chunlin Chen[1], Yu Cheng[3], Zhi Wang[12]

Welcome for communication
& cooperation!

Email: zicanhu@smail.nju.edu.cn
shilinzhang@smail.nju.edu.cn