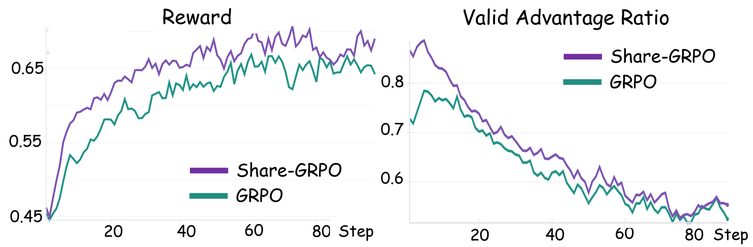
R1-ShareVL: Incentivizing Reasoning Capability of Multimodal Large Language Models via Share-GRPO

Huanjin Yao[2,3], Qixiang Yin[4], Jingyi Zhang[1], Min Yang[2], Yibo Wang[3], Wenhao Wu[5], Fei Su[4], Li Shen[6], Minghui Qiu[2], Dacheng Tao[1] Jiaxing Huang[1]

Paper GitHub WeChat

Challenge

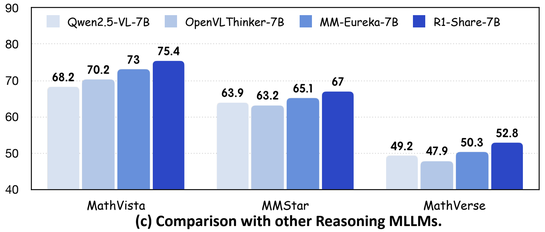
Reinforcement Learning has shown its promise in incentivizing the long-chain reasoning capability of MLLMs. However, applying GRPO to MLLMs often suffers from advantage vanishing, when all responses are correct or all are incorrect, the reward signal collapses to zero, thus affecting optimization effectiveness and overall training efficiency.



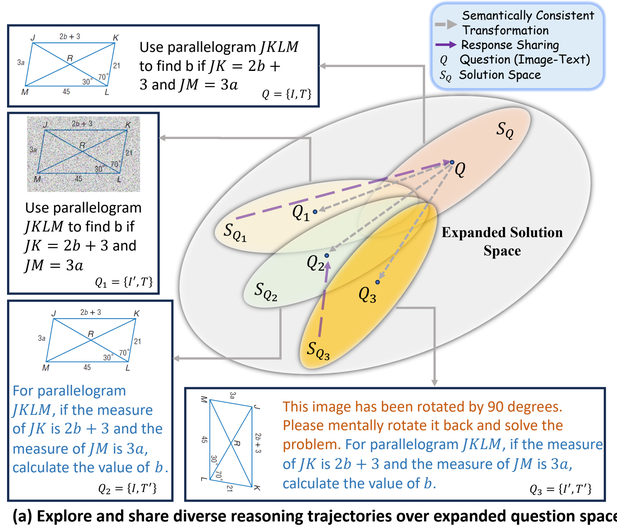
(b) Reward Curve (Left) and Valid Advantage Ratio Curve (Right)

Motivation

Share-GRPO tackles advantage vanishing by expanding the question space and sharing reasoning trajectories. It generates semantically equivalent question variants, explores diverse reasoning paths across them, and shares the explored trajectories and rewards to improve learning efficiency.



(c) Comparison with other Reasoning MLLMs.



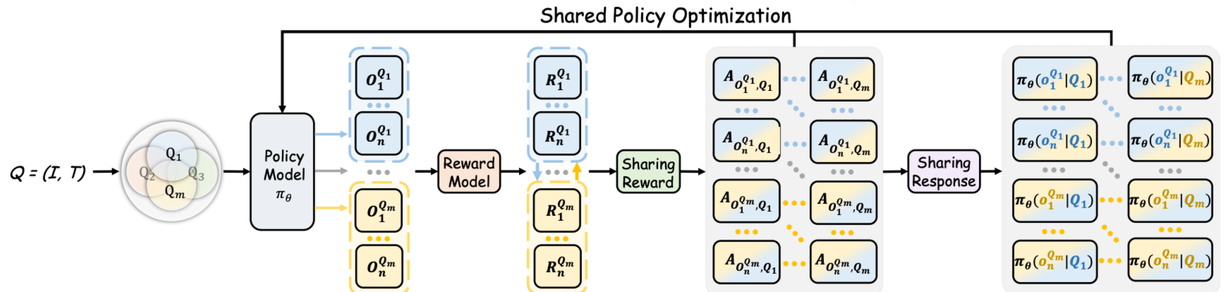
(a) Explore and share diverse reasoning trajectories over expanded question space

ShareGRPO

Question Space Expansion: Offline Textual Semantically Consistent Transformation, and Online Multimodal Semantically Consistent Transformation.

Shared Advantage Estimation: The outcome-level relative advantages are computed across the expanded and diverse trajectories.

Shared Policy Optimization: Optimize policy model by sharing diverse reasoning trajectories $\mathbf{O} = \{\{o_1^{Q_1}, ..., o_n^{Q_1}\}, ..., \{o_1^{Q_m}, ...o_n^{Q_m}\}\}$ across question variants $\mathbf{Q} = \{Q_1, Q_2, ..., Q_m\}$



Results

We compare R1-ShareVL which is trained by Share-GRPO without cold-start against SOTAs across multiple reasoning tasks, including both domain-specific and general-purpose tasks.

Model	MathVista	MMStar	MMMU	MathVerse	MathVision	AI2D	Avg.
GPT-4o[55]	63.8	65.1	70.7	50.8	30.4	84.9	60.9
Claude3.7-Sonnet[56]	66.8	–	71.8	52.0	41.3	–	–
Kimi1.5[1]	70.1	–	68.0	–	31.0	–	–
LLaVA-Reasoner-8B [57]	50.6	54.0	40.0	–	–	78.5	–
LLaVA-CoT-11B[26]	54.8	57.6	–	–	–	78.7	–
Mulberry-7B[27]	63.1	61.3	55.0	–	–	–	–
Qwen2.5-VL-7B [58] (Base Model)	68.2	63.9	58.6	49.2	25.1	83.9	58.1
X-REASONER-7B [59]	69.0	–	56.4	–	29.6	–	–
R1-Onevision-7B[37]	64.1	–	–	47.1	29.9	–	–
Vision-R1-7B[34]	73.5	64.3*	54.2*	52.4	29.4*	84.2*	59.7
OpenVLThinker-7B[39]	70.2	63.2	51.9	47.9	29.6	82.7	57.6
MM-Eureka-7B[5]	73.0	65.1*	55.3*	50.3	26.9	84.1*	59.1
ThinkLite-7B [60]	74.3	63.7	53.1	52.2	29.9	83.0	59.3
R1-ShareVL-7B	75.4	67.0	58.1	52.8	29.5	84.5	61.2
Scaling to Larger Models							
Qwen2.5-VL-32B [58] (Base Model)	74.7	69.5	70.0	49.9	38.4	84.6*	64.5
MM-Eureka-32B[5]	74.8	67.3*	64.6*	56.5	34.4	85.4*	63.8
R1-ShareVL-32B	77.6	70.2	70.1	59.0	40.3	86.2	67.2