R1-VL: Learning to Reason with Multimodal Large Language Models via Step-wise Group Relative Policy Optimization (Supplementary Materials)

We provide more experiment details in Section A (including datasets and implementation details), more experimental results in Section B, discussion about the related works in Section C and qualitative results in Section D.

A. Experiment details

A.1. Evaluation Benchmarks

Here we provide more details about the benchmarks used for evaluating the proposed StepGRPO:

MathVista [5]: MathVista is a benchmark that integrates challenges from a wide range of mathematical and visual tasks. It comprises 6,141 examples, sourced from 28 existing multimodal datasets related to mathematics, along with three newly created datasets: IQTest, FunctionQA, and PaperQA. Following prior studies [11], we adopt the mini version of MathVista, which consists of 1,000 questions, for evaluation.

MMStar [1]: MMStar is a vision-centric multimodal benchmark comprising 1,500 challenging samples. It is designed to assess six fundamental capabilities (i.e., coarse perception, fine-grained perception, instance reasoning, logical reasoning, science & technology and mathematics) across 18 specific dimensions, providing a well-balanced and refined evaluation of MLLMs' multimodal reasoning abilities.

Math-Vision [8]: MATH-Vision (MATH-V) benchmark is a carefully curated collection of 3,040 high-quality mathematical problems with visual contexts, sourced from real math competitions. It covers 16 distinct mathematical disciplines and is graded across five difficulty levels, offering a comprehensive and diverse challenge set for assessing the mathematical reasoning abilities of MLLMs.

ChartQA [6]: ChartQA is a large-scale benchmark designed to evaluate a model's ability to answer questions about charts, focusing on both visual and logical reasoning. It comprises 9,608 human-written questions and an additional 23,111 questions generated from human-written chart summaries. Following prior work, we evaluate StepGRPO on the test split of this dataset.

DynaMath [14]: DynaMath is a dynamic visual math benchmark designed for comprehensive evaluation of MLLMs. It consists of 501 high-quality, multi-topic seed questions, each represented as a Python program and covering a diverse range of mathematical topics, including Plane Geometry, Solid Geometry, Analytic Geometry, Algebra, Puzzle Tests, Graph Theory, Statistics, Scientific Figures, and Arithmetic. Each seed question is further expanded into 10 concrete variations, resulting in a total of 5,010 questions, ensuring a diverse and rigorous assessment of mathematical reasoning in multimodal models.

HallusionBench [4]: HallusionBench is a comprehensive benchmark designed to evaluate language hallucination and visual illusion in MLLMs. It comprises 346 images and 1,129 QA pairs, covering diverse topics such as mathematics, geography, sports, and statistics. The benchmark introduces controlled question groups to analyze models' response consistency and failure modes, along with novel metrics for quantifying hallucination and illusion.

MathVerse [12]: MathVerse is a comprehensive visual math benchmark designed for equitable and in-depth evaluation of MLLMs. It consists of 2,612 high-quality, multi-subject math problems with diagrams, sourced from publicly available datasets. Each problem is transformed by human annotators into six distinct versions: Text Dominant, Text Lite, Text Only, Vision Intensive, Vision Dominant, and Vision Only. In this study, we evaluate StepGRPO on five visual-related versions, excluding Text Only.

MME [3]: MME is a comprehensive benchmark designed to assess the performance of Multimodal Large Language Models (MLLMs). MME evaluates both perception and cognition abilities across 14 subtasks. To prevent data leakage, all instruction-answer pairs are manually crafted.

MMReason [10]: MMReason is a new benchmark designed to precisely and comprehensively evaluate the long-chain reasoning capabilities of MLLMs through diverse, open-ended, and challenging questions.

A.2. Prompt for key step pre-extraction

As discussed in Section Method in the main text, our step-wise reasoning accuracy provides additional rewards to reasoning path through a soft key-step matching mechanism.

```
PROMPT = """Given a reasoning process with structured reasoning steps
(### headers), extract the key elements from each step.

- For ### Image Description: Extract at most three key elements such as
named entities (objects, labels) and numerical values.
- For ### Rationales: Extract the core concepts needed to solve the
problem, such as mathematical properties or definitions.
- For ### Step X: Extract essential variables, equations, and
relationships that contribute to the final solution.
- Ignore ### Let's think step by step and ### ### The final answer is.

Retain only essential terms, numbers, and equations exactly as they
appear in the text. If a key element is a phrase (not a number or
equation), ensure it contains at most two words. Do not change or
rephrase any extracted key element. Only extract what is explicitly
stated in the original text.

Format your response as follows:
"Key Elements:\nImage Description: ["key1", "key2", ...]; Rationales:
["key1", "key2", ...]; Step 1: ["key1", "key2", ...], Step 2: ["key1",
"key2", ...]; ...".

{reasoning_path}"""
```

Figure 1. Prompt used for key-step extraction.

Specifically, we prompt GPT-4 to extract several key steps from the reasoning path for each question, where the prompt used for key-step extraction is shown in Fig. 1. To ensure efficient reward assignment, we ask GPT to refine the extracted steps by removing redundant content and retaining only the core few words necessary for reasoning.

B. More experimental results

Comparison with GRPO. From Table 1, we observe that applying GRPO directly to baseline models often results in performance degradation, primarily due to the sparse reward issue. In contrast, R1-VL with our proposed StepGRPO consistently improves vanilla GRPO by significant margins, which is largely attributed to that StepGRPO introduces step-wise reasoning accuracy and validity rewards.

C. Difference to other RL-based MLLM reasoning studies

We note that a few studies have also explored reinforcement learning for enhancing the reasoning capability of MLLMs [2, 13]. These methods typically adopt offline RL algorithms, such as Direct Preference Optimization (DPO), where models are optimized based on pre-collected reasoning preference data, i.e., reasoning paths leading to correct final answers are treated as preferred data, while reasoning paths leading to incorrect answers are used as rejected data. Different from these works, we propose StepGRPO, a novel online reinforcement learning framework that enables a MLLM to self-improve reasoning ability via simple, effective and dense step-wise rewarding. Our StepGRPO has two clear advantages: 1) StepGRPO operates in an online manner, allowing the model to generate and refine reasoning paths during the training process. This ensures that the learned reasoning strategies are up-to-date, rather than being constrained by a fixed dataset of reasoning trajectories; 2) Instead of solely relying on outcome-level reward signals, StepGRPO introduces step-wise reasoning rewards, including Step-wise Reasoning Accuracy Reward and Step-wise Reasoning Validity Reward. With the proposed step-wise reasoning rewards, StepGRPO enables to effectively mitigate the sparse reward issue for MLLMs without the need of process reward models and encourages more structured and logically consistent reasoning process.

D. Qualitative Results

Here we provide more qualitative results of the proposed StepGRPO. As shown in Fig. 2, our StepGRPO generates correct, logically consistent and structured reasoning paths.

E. Limitation

Our rule-based rewards are specifically designed for complex reasoning tasks that require step-by-step structured logical processes. As in Table 1 of the manuscript, they generalize well across a variety of tasks like math, medical and science QA, chart understanding, and improve more significantly on reasoning-heavy tasks (an average gain of 6% on four math benchmarks). On the other hand, our approach may be less effective for free-form tasks such as storytelling

Method	MathVista	MMStar	Math-V	ChartQA	DynaMath	HallBench	MathVerse	MME_{sum}	AVG
Qwen2-VL-2B [9]	43.0	48.0	12.4	73.5	24.9	41.7	19.7	1872	41.2
Qwen2-VL-2B-GRPO [7]	41.4	46.2	16.0	72.5	24.2	42.2	19.9	1930	41.4
R1-VL-2B (Ours)	52.1	49.8	17.1	75.2	29.4	44.0	26.2	2048	45.8
Qwen2-VL-7B [9]	58.2	60.7	16.3	83.0	42.1	50.6	32.5	2327	53.3
Qwen2-VL-7B-GRPO [7]	55.1	59.8	19.1	81.3	33.9	48.5	30.9	2335	51.4
R1-VL-7B (Ours)	63.5	60.0	24.7	83.9	45.2	54.7	40.0	2376	57.1

Table 1. Comparsion with GRPO.

or creative writing, where there is no single correct answer and no requirement for a structured intermediate reasoning process.

References

[1] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 1

[2] Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. *arXiv preprint arXiv:2411.14432*, 2024. 2

[3] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 1

[4] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*, 2023. 1

[5] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 1

[6] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 1

[7] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 3

[8] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2025. 1

[9] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

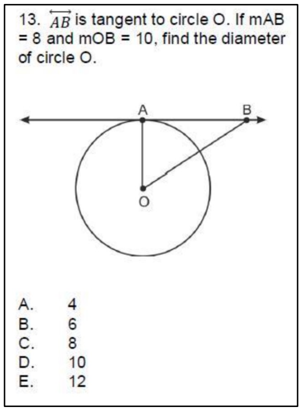
[10] Huanjin Yao, Jiaxing Huang, Yawen Qiu, Michael K Chen, Wenzheng Liu, Wei Zhang, Wenjie Zeng, Xikun Zhang, Jingyi Zhang, Yuxin Song, et al. Mmreason: An open-ended multi-modal multi-step reasoning benchmark for mllms toward agi. *arXiv preprint arXiv:2506.23563*, 2025. 1

[11] Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*, 2024. 1

[12] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024. 1

[13] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*, 2024. 2

[14] Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836*, 2024. 1



Question: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.

Correct answer: (E) 12

StepGRPO:

Image Description: The image shows a circle with center O. A tangent line AB touches the circle at point A. The line OB is a radius of the circle, and the length of AB is given as 8. The length of OB is given as 10.

Rationales: The problem involves a tangent to a circle and requires finding the diameter. The tangent line is perpendicular to the radius at the point of tangency, forming a right triangle with the radius and the tangent segment.

Let's think step by step.

Step 1: Identify the right triangle formed by the tangent AB, radius OB, and the line segment OA. Since AB is tangent to the circle at A, angle OAB is 90 degrees.

Step 2: Use the Pythagorean theorem in triangle OAB. The sides are OA (radius), OB (hypotenuse), and AB (tangent).

Step 3: Apply the Pythagorean theorem: \(OA^2 + AB^2 = OB^2 \).

Step 4: Substitute the known values: \(OA^2 + 8^2 = 10^2 \).

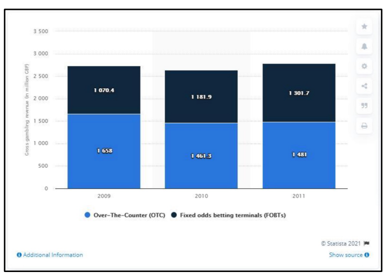
Step 5: Calculate: \(OA^2 + 64 = 100 \).

Step 6: Solve for \(OA^2 \): \(OA^2 = 100 - 64 = 36 \).

Step 7: Find OA: \(OA = \sqrt{36} = 6 \).

Step 8: Since OA is the radius, the diameter is twice the radius: \(2 \times 6 = 12 \).

The final answer is: E



Question: What is the sum of highest value and lowest value of navy blue bar?

Correct answer: 2372.1

StepGRPO:

Image Description: The image is a bar chart showing gross gambling revenue in million GBP for the years 2009, 2010, and 2011. Each year has two bars: a blue bar labeled "Over-The-Counter (OTC)" and a navy blue bar labeled "Fixed odds betting terminals (FOBTs)." The values for the navy blue bars are 1,070.4 for 2009, 1,181.9 for 2010, and 1,301.7 for 2011.

Rationales: The question asks for the sum of the highest and lowest values of the navy blue bars, which represent the Fixed odds betting terminals (FOBTs) for each year.

Let's think step by step.

Step 1: Identify the values of the navy blue bars for each year:

- 2009: 1,070.4

- 2010: 1,181.9

- 2011: 1,301.7

Step 2: Determine the highest and lowest values among these:

- Highest value: 1,301.7 (2011)

- Lowest value: 1,070.4 (2009)

Step 3: Calculate the sum of the highest and lowest values:

- Sum = 1,301.7 + 1,070.4

The final answer is: 2372.1

Figure 2. Our StepGRPO generates correct, logically consistent and structured reasoning paths.