All Roads Lead to Rome: Incentivizing Divergent Thinking in Vision-Language Models

Supplementary Material

A. Experimental Details

Here we provide detailed descriptions of benchmarks, baseline models, and training configurations for reproducibility.

Benchmarks. In the main paper, we provide a high-level overview of the selected benchmarks. Here we present additional details, including dataset types, sizes, and data splits.

1) **MathVerse** [26] is a curated benchmark consisting of 2,612 mathematically focused questions, each featuring varying degrees of multimodal context. Consistent with prior studies [8, 23], we conduct evaluation on the Test Mini split under the Vision Only setting, which contains approximately 700 samples.

2) **MathVista** [12] is a large-scale benchmark designed to assess diverse mathematical reasoning skills, including logic puzzles, algebraic manipulation, and scientific interpretation. It consists of 6,141 annotated instances. In our experiments, we evaluate models on the Test Mini split, which includes around 1,000 samples.

3) **MathVision** [18] is a challenging benchmark consisting of 3,040 carefully curated mathematical problems drawn from real-world math competitions. Covering 16 subject areas and five difficulty tiers, it offers a broad and rigorous testbed for evaluating the reasoning capabilities of VLMs. Our evaluation is performed on the complete test set.

4) **LogicVista** [21] is designed to evaluate core logical reasoning abilities in VLMs, spanning various categories such as spatial, deductive, inductive, numerical, and mechanical reasoning. The benchmark includes 448 visual multiple-choice questions. In our study, we perform evaluation on the full test split.

5) **WeMath** [15] is a visual mathematics benchmark containing approximately 6.5K problems organized into 67 hierarchical knowledge categories spanning five levels of conceptual granularity. For evaluation, we adopt the Test Mini split, which includes around 1,740 samples, and use the strict score as the principal evaluation metric.

6) **Geometry3k** [11] is a benchmark comprising 3,002 geometry questions annotated in formal mathematical language, designed to assess abstract reasoning and symbolic problem-solving grounded in axiomatic principles. For evaluation, we merge the validation and test sets, yielding approximately 900 examples.

7) **MMStar** [3] is a multimodal benchmark constructed to ensure strong visual grounding in every example, requiring advanced reasoning that cannot be solved without visual input. It includes 1,500 samples for offline evaluation. In our experiments, we evaluate using the full test set.

8) **HallusionBench** [7] is curated to assess the capabilities of advanced VLMs in performing fine-grained visual interpretation and understanding. The benchmark comprises 346 images paired with 1,129 expert-annotated questions. In this work, we evaluate models on the complete test split.

9) **MMVet** [25] consists of 218 examples and targets six fundamental vision-language capabilities, emphasizing their joint evaluation to assess the interplay between skills. The benchmark is intended to measure the overall proficiency of generalist VLMs. In our experiments, we evaluate on the full test split.

Baseline Models. Here we provide detailed descriptions of the experimental settings used for baseline models.

1) **R1-OneVision** [23] is available in both 3B and 7B model variants, trained using a two-stage process comprising SFT followed by RL, with an emphasis on mathematical reasoning. In our experiments, we utilize the publicly released 7B checkpoint built upon the Qwen2.5-VL-7B architecture.

2) **VLAA-Thinker** [2] is an RL–only model trained on a curated and challenging multimodal dataset. It is among the first to show that RL can surpass SFT in vision-language tasks. For our experiments, we adopt VLAA-Thinker-7B variant, which is built upon the Qwen2.5-VL-7B backbone.

3) **Vision-R1** [8] is a large-scale reasoning model trained with a significantly larger data volume, approximately five times that used in our setting. In our experiments, we evaluate the 7B variant, which represents the strongest publicly reported checkpoint according to the paper.

4) **VLM-R1** [16] is a general-purpose VLM designed for broad visual tasks including classification and detection through vision-centric reward shaping. In our experiments, we use VLM-R1-3B-Math, a publicly available checkpoint enhanced for mathematical reasoning.

5) **LMM-R1** [14] leverages RL on both text-only and multimodal data, progressively enhancing the model's reasoning capabilities through a multi-stage training framework. In our experiments, we use LMM-R1-MGT, which ranks among the top-performing models within the small-scale VLM category.

Training Configuration. In Section 5, we briefly outline the training setup. Here, we provide a more detailed description. By default, all models are trained on ViRL39K, a carefully curated dataset constructed by consolidating and refining samples from several prior benchmarks, including MM-Eureka [13], MV-Math [19], and M3CoT [4]. We reserve a hold-out set of 1000 examples for validation, using the remainder for training. All models are trained for 2 epochs with a learning rate of $1e^{-}6$. For ablation studies, we train on 5000 examples randomly sampled from the full training set for computational efficiency. During rollout, we generate $N = 15$ responses per example using a

sampling temperature of 1.0, and partition them into $K = 3$ groups, with a minimum group size $G_{\min} = 3$. We set the load-balance weight exponent $\beta = 1$, and annealed the diversity reward weight from an initial value $\lambda_{\max} = 0.4$ to a final value $\lambda_{\min} = 0.1$. All experiments are conducted using 8 NVIDIA A100 GPUs. Training and evaluation scripts are performed within the VeRL [17] and VLMEvalKit [6] frameworks, respectively.

B. More Ablation Studies

The load-balance weight exponent β. In the main text, we introduce the load-balance weight w_k to control each group's contribution to the overall optimization objective, mitigating the risk of larger groups dominating the learning process. Specifically, $w_k = \left(\frac{N}{K|G_k|}\right)^{\beta}$, where β determines the sensitivity of the weight to varying group size. This formulation assigns higher weights to smaller groups and vice versa. Here, we further investigate the effect of varying β on model performance. For computational efficiency, we conduct this analysis using 5000 examples randomly sampled from the full training set.

β	MathVerse	MathVista	MathVision	MMStar	HallBench	Average
0.0	47.9	70.6	24.8	65.2	55.3	52.8
0.5	50.1	72.8	27.4	65.7	56.0	54.4
1.0	51.2	**74.1**	**29.3**	**65.8**	56.5	**55.4**
1.5	**51.4**	73.7	28.9	65.7	**56.7**	55.3
2.0	51.0	73.5	28.3	65.2	56.1	54.8

Table 5. The benchmark results varying load-balance exponent β

As shown in Table 5, when β increases from 0.0, where all groups are equally weighted, to 2.0, which imposes a strong penalty on imbalanced groups, the model's performance peaks around $\beta = 1.0$ or $\beta = 1.5$. A very small β leads the model to collapse from a multimodal to a unimodal optimization pattern, ultimately degrading to GRPO, while an excessively large β enforces overly uniform group sizes, hindering effective exploitation in later training stages.

The minimum group size $G_{\min}$. In the main text, we divide N responses into K groups of varying size depending on their reasoning embeddings. To ensure the reliability of advantage estimation within each group, we enforce a minimum group size $G_{\min}$. Here, we investigate how different choices of $G_{\min}$ affect the overall performance of the model.

As shown in Table 6, the model achieves optimal performance when the minimum group size $G_{\min} = 3$, which serves as our default configuration. Reducing $G_{\min}$ to 2 leads to a notable drop in accuracy, suggesting that groups with insufficient response samples fail to produce reliable advantage estimates. Conversely, increasing $G_{\min}$ to 4 or 5 results in a slight decline in performance, likely due to the overly rigid enforcement of group balance, which constrains training flexibility and impedes the model's ability to exploit and refine particularly effective solution strategies.

$G_{\min}$	MathVerse	MathVista	MathVision	MMStar	HallBench	Average
2	50.5	73.3	28.1	65.3	56.2	54.7
3	**51.2**	74.1	**29.3**	**65.8**	**56.5**	**55.4**
4	51.0	**74.7**	28.5	65.5	56.2	55.2
5	50.6	73.9	28.2	65.1	56.3	54.8

Table 6. The impact of minimum group size $G_{\min}$.

The rollout number of responses N. In contrast to GRPO, which operates on a single group of responses, our approach partitions N responses into multiple groups, necessitating a larger N to ensure adequate group sizes. A natural question arises: Does MUPO's performance gain stem primarily from its encouragement of divergent thinking, or is it merely a result of increased sampling? To disentangle these factors, we conduct a comparison between MUPO and GRPO under the same total number of responses N, as shown in Table 7.

Method	MathVerse	MathVista	MathVision	MMStar	HallBench	Average
$\text{GRPO}_{\text{N=5}}$	46.4	69.1	23.7	64.6	54.3	51.6
$\text{GRPO}_{\text{N=8}}$	46.8	69.6	23.6	65.1	54.6	52.0
$\text{GRPO}_{\text{N=12}}$	46.5	69.4	23.8	64.9	54.1	51.7
$\text{GRPO}_{\text{N=15}}$	46.9	69.1	24.1	64.8	54.7	51.9
$\text{MUPO}_{\text{N=15}}$	**51.2**	**74.1**	**29.3**	**65.8**	**56.5**	**55.4**

Table 7. The effect of rollout number N on model performance.

It can be seen that MUPO significantly outperforms GRPO even when both algorithms use the same number of sampled responses $N = 15$. Moreover, we observe that the model performance is insensitive to the choice of N, where accuracy remains stable across N ranging from 8 to 15. These findings indicate that MUPO's superiority does not stem from an increased number of sampled sequences, but rather from the integration of reasoning diversity, enabling the discovery of more effective solution strategies.

C. Choice of Embedding Models

In the main text, we compute the diversity reward by extracting reasoning embeddings using Qwen3-Embedding-0.6B and measuring pairwise cosine distances. This raises an important question: Is the embedding space derived from such a lightweight model sufficiently reliable for capturing semantic similarity in reasoning? Or is there a better embedding model to handle this? To answer the question, we explore alternative embedding models with larger scales and different architectures. Specifically, we experi-

ment with Qwen3-Embedding models [27] at the 4B and 8B parameter scales, as well as LLaMA-Embed-8B [10]. Additionally, we consider an ensembling strategy, where reasoning distances are computed by averaging outputs from multiple embedding models to provide a more robust estimation.

Model	MathVerse	MathVista	MathVision	MMStar	HallBench	Average
$\text{Qwen}_{0.6B}$	51.2	**74.1**	29.3	65.8	56.5	55.4
Qwen_{4B}	51.0	73.9	29.1	**66.6**	57.3	55.6
Qwen_{8B}	**51.6**	74.1	28.7	65.4	56.7	55.3
LLaMA_{8B}	50.6	73.6	28.9	65.1	55.8	54.8
Ensemble	51.5	73.9	**29.6**	66.5	**57.5**	**55.8**

Table 8. The impact of choices of different embedding models.

As shown in Table 8, different scales of the Qwen3-Embedding model have marginal impact on performance (within $\pm 0.3\%$), while LLaMA-Embed-8B performs noticeably worse than the Qwen series. This suggests that Qwen3-Embedding-0.6B offers the best trade-off between effectiveness and efficiency. Furthermore, the ensemble approach yields slightly better results than any single model, indicating complementary effects among different embedding models.

D. More Test-time Scaling Results

In the main text, we introduce acc@4 as a metric that is counted as correct if at least one of four generated responses produces the right answer, serving to validate the strong test-time scaling capability of our model. However, in practical applications, even when a correct answer exists among multiple responses, an external verifier is still required to identify it. Consequently, here we further explore a more commonly adopted test-time scaling approach, self-consistency, which aggregates multiple responses through majority voting without the need for a verifier. Specifically, we employ maj@16, where 16 responses are sampled for each question, and the most frequently predicted answer is selected as the final output.

Model	MathVerse	MathVista	MathVision	MMStar	HallBench	Average
QwenVL	45.3	66.7	27.8	63.6	56.5	52.0
InternVL	37.9	69.6	27.7	66.4	52.2	50.7
R1-OV	48.2	66.3	<u>30.9</u>	65.7	54.5	53.1
VLAA	49.8	68.7	28.1	<u>68.0</u>	55.4	54.0
V-R1	<u>53.8</u>	<u>74.3</u>	30.7	67.1	<u>57.0</u>	<u>56.6</u>
MUPO	**57.1**	**76.5**	**35.8**	**71.8**	**59.9**	**60.2**

Table 9. The maj@16 scores of MUPO and existing baselines.

As shown in Table 9, our model outperforms the previous best results by an average gain of 3.6% ($56.6\% \rightarrow 60.2\%$) on the maj@16 metric. This improvement indicates that divergent thinking substantially increases the frequency of correct answers among the generated responses, thereby enabling majority voting to effectively strengthen the model's test-time scaling capability, consistent with the conclusions presented in the main text.

E. MUPO as a Plug-and-Play

In the main text, we primarily compare MUPO with GRPO, as MUPO can be viewed as a direct extension of GRPO from single-group to multi-group optimization to encourage divergent thinking in VLMs. Notably, the strategy of multi-group advantage estimation and diversity reward design is inherently modular and can serve as a plug-and-play enhancement to a wide range of policy optimization algorithms, since divergent training is orthogonal to most existing optimization strategies. To validate this, in addition to GRPO, we also consider two widely adopted algorithms including DAPO [24] and GSPO [28], which aim to improve learning efficiency from the perspectives of sampling and importance ratio, respectively. We denote the resulting variants as MUPO_{G}, MUPO_{D}, and MUPO_{S}, corresponding to the integration of MUPO with GRPO, DAPO, and GSPO.

Method	MathVerse	MathVista	MathVision	MMStar	HallBench	Average
GRPO	46.9	69.1	24.1	64.8	54.7	51.9
MUPO_{G}	51.2	74.1	29.3	65.8	56.5	55.4
DAPO	48.5	70.3	25.7	65.5	55.2	53.1
MUPO_{D}	**52.7**	**75.2**	<u>31.0</u>	**66.9**	<u>56.7</u>	<u>56.5</u>
GSPO	47.8	71.9	27.4	64.9	56.4	53.7
MUPO_{S}	<u>52.5</u>	<u>74.9</u>	**32.6**	<u>66.8</u>	**57.6**	**56.9**

Table 10. MUPO as a plug-and-play for various policy algorithms.

As shown in Table 10, incorporating MUPO consistently enhances the performance across different policy optimization algorithms. Specifically, we observe average accuracy gains of 3.5%, 3.4% and 3.2% when applied to GRPO, DAPO, and GSPO, respectively. These results reveal a significant bottleneck in existing methods due to their convergent training, which can be mitigated through our proposed multi-group optimization. This also highlights the complementary nature of MUPO with these strong baselines.

F. MUPO on Other Models

In the main text, our experiments primarily utilize Qwen2.5-VL models [1] at the 3B and 7B scales as base models, which may raise the question of whether MUPO can extend to other architectures. To assess the generalizability of our approach, we further employ MUPO on the latest model, Qwen3-VL-8B [22].

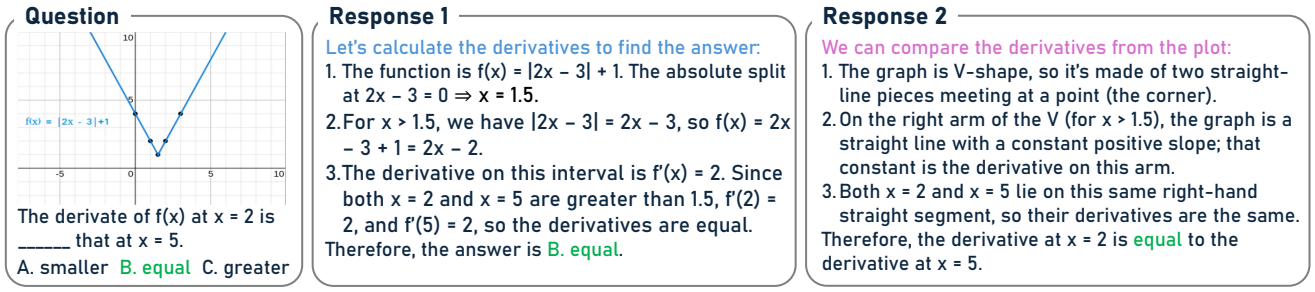

Figure 10. More qualitative examples of MUPO. The example is selected from the MathVista [12], where we sample multiple responses.

Model	MathVerse	MathVista	MathVision	MMStar	HallBench	Average
Qwen2.5-7B	40.7	62.3	23.2	59.2	50.0	47.1
InternVL-8B	34.5	68.2	25.6	63.2	49.0	48.1
R1-OV-7B	46.4	64.1	29.9	64.7	52.5	51.5
VLAA-7B	48.2	68.0	26.4	66.1	54.7	52.7
V-R1-7B	52.4	73.5	28.2	66.3	55.4	55.2
Qwen2.5$_{\text{MUPO}}$	53.9	77.9	31.3	68.7	57.5	57.9
Qwen3-8B	58.8	74.8	51.5	67.3	60.2	62.5
Qwen3$_{\text{GRPO}}$	63.5	79.0	54.3	70.3	62.2	65.8
Qwen3$_{\text{MUPO}}$	**69.4**	**83.1**	**58.4**	**71.6**	**65.8**	**69.7**

Table 11. The evaluation results of MUPO on Qwen3-VL-8B.

Model	Mathematics		General		Average	
	Acc@1	Acc@4	Acc@1	Acc@4	Acc@1	Acc@4
Qwen2.5-VL-7B	40.1	56.5	58.0	71.5	46.0	61.5
+ GRPO	46.5	51.7	62.7	66.5	52.2	58.4
+ Entropy Reg.	47.3	53.2	63.2	67.8	53.4	60.2
MUPO-Thinker-7B	**51.6**	**58.8**	**65.6**	**72.4**	**56.3**	**63.3**

Table 12. Comparison between MUPO and entropy regularization.

As shown in Table 11, the Qwen3-VL base model and its RL-enhanced variant significantly outperform their Qwen2.5-VL counterparts across multiple established benchmarks. Despite the already strong performance of Qwen3-VL, integrating MUPO still yields substantial gains, improving average accuracy by 7.2% ($62.5\% \rightarrow 69.7\%$) over the base model and by 3.9% ($65.8\% \rightarrow 69.7\%$) over the GRPO variant. These results underscore the complementary benefits of MUPO, even when applied to more advanced and competitive models.

G. Entropy Regularization

The effectiveness of MUPO in enhancing model performance primarily stems from its ability to foster divergent thinking, enabling models to explore multiple reasoning branches during training rather than prematurely converging to a narrow solution space. This motivation resembles a conventional approach to promote exploration in RL, *i.e.*, entropy regularization. Recent studies [5, 9, 20] have explored entropy-based techniques to encourage exploration and improve training efficiency. However, it remains unclear whether this naive remedy enables models to discover globally optimal solutions and learning distinct reasoning modes. To further investigate this, we compare MUPO with entropy regularization in terms of acc@1 and acc@4 scores. For entropy regularization, we adopt the widely used average entropy maximization, which has been implemented within the VeRL framework [17].

As shown in Table 12, even when GRPO is augmented with entropy regularization, its performance remains notably lower than that of MUPO, by 2.9% ($53.4\% \rightarrow 56.3\%$) on acc@1 and 3.1% ($60.2\% \rightarrow 63.3\%$) on acc@4, respectively. This indicates that simply maximizing entropy provides only marginal benefits: while it can encourage exploration, it fails to guide models toward learning effective reasoning strategies. In contrast, MUPO achieves this by promoting both breadth and depth in exploration.

H. More Examples

For completeness, we present additional responses generated by MUPO-Thinker-7B in Fig. 10. The task involves comparing the derivative values of a function at two different points. Notably, MUPO-Thinker-7B demonstrates the ability to approach the problem via two distinct pathways: it can either compute and compare the derivatives analytically or infer their relative magnitudes directly from the provided graph. This dual capability underscores the model's flexibility and diversity in problem-solving strategies, embodying the principle that "All Roads Lead to Rome".

I. Limitation Analysis

Computational Cost. To promote divergent thinking, MUPO deviates from GRPO's single-group design by partitioning the sampled responses into multiple groups, thereby encouraging exploration across diverse reasoning trajectories. This multi-group structure typically desires a larger number of responses, resulting in increased computational overhead during the rollout phase. In practice, we moderately relax the strict on-policy requirement by occasionally incorporating responses generated in earlier steps, which

can be viewed as a trick similar to experience replay. This strategy allows MUPO to outperform GRPO significantly while operating under a comparable computational budget.

Hyperparameter Sensitivity. Compared to GRPO, MUPO introduces additional hyperparameters including the number of groups K, minimum group size $G_{\min}$, load-balance exponent β, and the initial and final values of diversity reward weight $\lambda_{\max}$ and $\lambda_{\min}$. These parameters play a critical role in shaping the training dynamics of MUPO and may exhibit optimal configurations depending on the characteristics of the target task, as exemplified in Table 4 of the main paper. Consequently, enabling dynamic, task-dependent adjustment of key hyperparameters, *e.g.*, the number of groups K, presents a promising direction for future research.

J. Societal Broader Impact

Our work presents a positive societal impact. Beyond achieving higher accuracy, MUPO holds significant practical value across a variety of real-world applications. For instance, in AI for education, students often rely on the reasoning provided by VLMs to aid their learning. In such scenarios, diverse solution strategies are particularly beneficial, as they expose learners to multiple lines of thought, fostering deeper understanding and critical thinking. Similarly, in open-ended tasks such as creative writing, diversity in generation is considered a key metric for evaluating models' generalization and versatility. Furthermore, the ability to reason from multiple perspectives enhances interpretability and fosters trust in model outputs by allowing users to better assess the reliability of the generated answers.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3

[2] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*, 2025. 1

[3] 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? *NeurIPS*, 37:27056–27087, 2024. 1

[4] Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *ACL*, 2024. 1

[5] Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*, 2025. 4

[6] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *ACM MM*, pages 11198–11201, 2024. 2

[7] 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. In *CVPR*, pages 14375–14385, 2024. 1

[8] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025. 1

[9] Yuhua Jiang, Jiawei Huang, Yufeng Yuan, Xin Mao, Yu Yue, Qianchuan Zhao, and Lin Yan. Risk-sensitive rl for alleviating exploration dilemmas in large language models. *arXiv preprint arXiv:2509.24261*, 2025. 4

[10] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models, 2024. 3

[11] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *ACL*, 2021. 1

[12] 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. *ICLR*, 2023. 1, 4

[13] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025. 1

[14] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025. 1

[15] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *ACL*, 2024. 1

[16] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025. 1

[17] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and

Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297, 2025. 2, 4

[18] 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. *NeurIPS*, 37:95095–95169, 2024. 1

[19] Peijie Wang, Zhong-Zhi Li, Fei Yin, Dekang Ran, and Cheng-Lin Liu. Mv-math: Evaluating multimodal math reasoning in multi-visual contexts. In *CVPR*, pages 19541–19551, 2025. 1

[20] Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025. 4

[21] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024. 1

[22] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 3

[23] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025. 1

[24] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. 3

[25] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *ICML*, 2023. 1

[26] 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 *ECCV*, pages 169–186. Springer, 2024. 1

[27] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. 3

[28] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025. 3