
OpenVLThinker: Complex Vision-Language Reasoning via Iterative SFT-RL Cycles

Yihe Deng, Hritik Bansal, Fan Yin
Nanyun Peng, Wei Wang, Kai-Wei Chang

University of California, Los Angeles

Abstract

We introduce *OpenVLThinker*, one of the first open-source large vision-language models (LVLMs) to exhibit sophisticated chain-of-thought reasoning, achieving notable performance gains on challenging visual reasoning tasks. While text-based reasoning models show promising results in text-only tasks, distilling their reasoning into LVLMs via supervised fine-tuning (SFT) often results in performance degradation due to imprecise visual grounding. Conversely, purely reinforcement learning (RL)-based methods face a large search space, hindering the emergence of reflective behaviors in smaller models. Surprisingly, alternating between SFT and RL ultimately results in significant performance improvements after a few iterations. Our analysis reveals that the base model rarely exhibits reasoning behaviors initially, but SFT effectively surfaces these latent actions and narrows the RL search space, accelerating the development of reasoning capabilities. Each subsequent RL stage further refines the model’s reasoning skills, producing higher-quality SFT data for continued self-improvement. OpenVLThinker-7B consistently advances performance across six benchmarks demanding mathematical and general reasoning.

1 Introduction

Proprietary large language models (LLMs), notably OpenAI’s o-series [29] and Google’s Gemini-2.5 Pro [17], have demonstrated impressive multi-step reasoning abilities of planning, reflection, and verification. Recent open-weight models [27, 46, 51, 80, 81] (e.g., DeepSeek-R1 [19] and smaller LLMs like S1 [48] and QwQ-32B [61]) show that reinforcement learning (RL) with verifiable rewards effectively reproduces these advanced capabilities, significantly boosting performance on challenging mathematical and logical tasks. Unlike text-only LLMs, it remains unclear whether open-source large vision-language models (LVLMs) can effectively adopt similar sophisticated reasoning strategies.

Moreover, it is known that reasoning capabilities can generally be distilled from larger LLMs to smaller ones through supervised fine-tuning (SFT) on chain-of-thought demonstrations [31, 35] for text-only tasks. This recipe has been recently applied in distills demonstrations from DeepSeek-R1 [48, 60, 75]. However, adapting this method to LVLMs does not work. Proprietary LVLMs, such as OpenAI’s o1/o3, do not expose their internal reasoning paths, making their outputs unsuitable for distillation. Therefore, most recent attempts are focusing on improving LVLMs through distillation from text-only R1 reasoning models. Unfortunately, our experiments show that naively fine-tuning LVLMs on reasoning paths generated from text-based DeepSeek-R1 with image captions leads to a non-trivial performance drop (see Figure 3), primarily due to a lack of precise visual grounding. Similar observations can be found in [6, 73].

In this paper, we present OpenVLThinker-7B, one of the **first** open-weight LVLMs that exhibit complex reasoning capabilities in complex vision-language tasks. Specifically, it is trained by iterating between the following two steps:

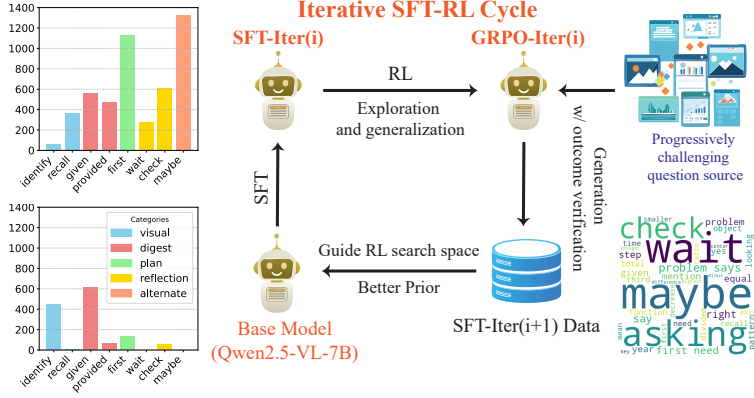


Figure 1: Illustration of OpenVLThinker-7B’s training process. We iteratively apply SFT and GRPO to refine the LVLM using reasoning data generated from previous iterations.

1. *Lightweight SFT*. In the first iterations, we distill CoTs using a text-only Deep-Seek R1 given the task question and the corresponding generated image caption. These CoT traces provide demonstrations of reasoning actions, although they do not immediately improve LVLM’s accuracy. For later iterations, we use the LVLM from the previous iteration to produce CoTs on 3,000 data points. This small dataset is sufficient to progressively enhance the model’s reasoning depth.
2. *Curriculum RL*. In subsequent iterations, we further enhance the LVLM’s reasoning through RL exploration with *Group Relative Policy Optimization* (GRPO) [54], which splits training into two rounds to form a smooth curriculum.

We found that while the initial step of SFT leads to a performance drop, iteratively alternating between SFT and RL eventually gradually yields a significant performance gain on both reasoning depth and answer accuracy (Figure 1). Our further analysis shows that the inference-time reasoning behaviors are often triggered by specific tokens (e.g., "wait"). SFT serves as an inductive prior that highlights these reasoning actions, where tokens such as “wait” trigger the model’s planning, reflection, and verification behaviors. Without this SFT step, launching RL from scratch forces the model to search through a prohibitively large space, making reflective behaviors slow to emerge, if they emerge at all. On the other hand, RL plays the critical role in learning the reasoning behaviors, generalizing from training data, and offering a better foundation for the next SFT iteration. The iterative cycle between SFT and RL collaboratively optimizes LVLM’s performance. We highlight our contributions as follows:

- We introduce **OpenVLThinker-7B**, one of the first open-source LVLMs to demonstrate reliable self-reflection, planning, and correction in visual contexts.
- We present a simple yet effective iterative SFT-RL loop that enables R1-style reasoning into multimodal domains and steadily self-improves without requiring massive datasets.
- On six challenging benchmarks, including MathVista and MathVerse, OpenVLThinker presents remarkable improvements while reducing hallucination on HallusionBench.

2 OpenVLThinker: Iterative Self-improvement on Curriculum Data

2.1 The Role of SFT and RL

The initial SFT data. The standard distillation approach used for text-only reasoning cannot be directly applied because the R1 model does not support visual input, and other proprietary LVLMs, such as OpenAI’s o1/o3, do not expose their internal reasoning paths. To learn reasoning behaviors from R1, we instead use the target model as a captioning model, prompting it to generate detailed textual descriptions for each image. Subsequently, these captions serve as proxies for the images when input into a text-based R1 reasoning model, *QwQ-32B* [61], which then generates k candidate reasoning chains. Among these candidates, we select the shortest reasoning chain that correctly arrives at the final answer to avoid excessive reasoning length after SFT.

Impact of SFT and RL on Model Reasoning Actions. Complex reasoning behaviors in LLMs have been described using various terms, including long CoT [76] and aha moments [19]. At their core, these behaviors reflect autonomous planning, reflection, and verification steps that occur during inference. We refer to them as *inference-time actions*, which are often triggered by specific tokens

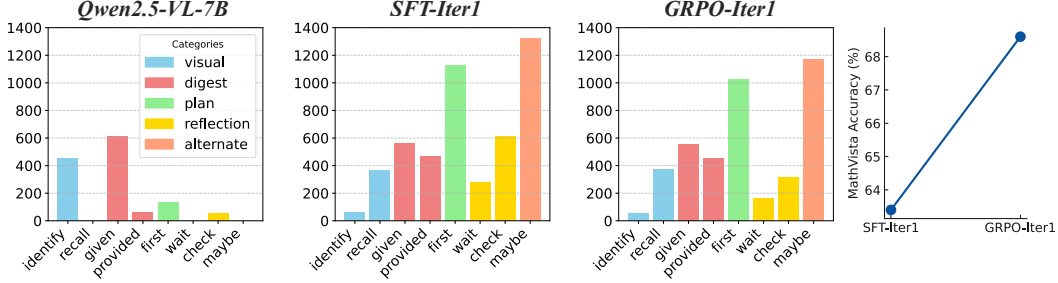


Figure 2: Occurrences of reasoning keywords when solving MathVista with the base model, SFT-Iter1 model, and GRPO-Iter1 model. The most significant distribution shift occurs after SFT, while the scale remains largely unchanged after GRPO, despite notable performance improvements.

such as “wait”. To examine how SFT and RL influence these reasoning actions, we identify eight representative keywords corresponding to perception, question comprehension, planning, reflection, and seeking alternatives. As illustrated in Figure 2, the base model seldom exhibits planning, reflection, or alternative-solution actions. However, SFT guided by text-based RL-like reasoning traces effectively surfaces these behaviors. As shown in the third and fourth subplots of Figure 2, subsequent GRPO-based RL training following SFT-Iter1 substantially enhances model performance on MathVista by 5.2%, yet largely maintains the initial reasoning action distribution, with minor refinements such as reduced repetitive reflections.

Conversely, direct RL training without prior SFT struggles to efficiently induce reasoning behaviors, exemplified by the absence of reflection keywords (e.g., “wait”) even after an equivalent training volume. Concurrent research by [65], which solely relies on RL, addresses this by explicitly appending relevant keywords during training rollouts. These observations support our argument that SFT plays a critical role in highlighting desirable reasoning actions, providing an efficient and effective foundation for RL to build upon. In contrast, RL primarily serves to further refine and enhance performance.

2.2 Iterative Improvement

The model obtained after the first iteration (GRPO-Iter1) demonstrates enhanced complex reasoning capabilities and improved reliability in processing visual inputs compared to methods based on image-to-text conversion. This advancement positions GRPO-Iter1 as an effective source for generating higher-quality reasoning demonstrations. Consequently, we propose an iterative self-improvement strategy, inspired by established methodologies such as iterative SFT in ReST-EM [57] and iterative direct preference optimization (DPO) schemes [50, 78], both of which have shown substantial effectiveness in iterative training processes and fall under the Expectation-Maximization framework [57].

Specifically, in each iteration, we sample a new set of enhanced reasoning traces using the model trained in the preceding iteration. These refined demonstrations are then utilized to retrain the base model¹, thereby progressively elevating its reasoning performance. The overall iterative pipeline is illustrated in Figure 1, and the consistent incremental performance gains achieved through successive iterations are depicted in Figure 3. We defer further method details to Appendix C

3 Experiments

Training setup. We take Qwen2.5-VL-7B [3] as the base model and perform three iterations of the SFT-RL cycle as illustrated in Section 2, applying full fine-tuning for both SFT and RL. Our training

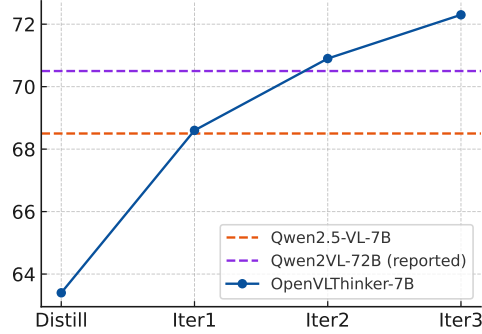


Figure 3: Iterative performance improvement of our model on MathVista. We note that $Iter(i)$ is always fine-tuned from the base model Qwen2.5-VL-7B, with its training data generated from $Iter(i-1)$.

¹To maintain stability, we retrain the model from scratch at each iteration with the newly generated dataset, as similar to some iterative approaches in text-only domain [23, 57].

Table 1: Evaluation results across visual math reasoning benchmarks (MathVista, MathVerse, MathVision), general visual reasoning benchmarks (MMMU-Pro, EMMA), and perception (HallusionBench). *Performance of the base model Qwen2.5-VL-7B and concurrent reasoning models are evaluated by us under the same setting and hardware as OpenVLThinker.

Model	Data	Math Reasoning			General Reasoning		Visual	Avg
		Math-Vista	Math-Verse	Math-Vision	MMMU-Pro	EMMA	Hallu-Bench	
Proprietary Model								
GPT-4o	-	63.8	50.2	30.4	51.9	32.7	55.0	47.3
Claude-3.5-Sonnet	-	67.7	47.8	33.5	51.5	35.1	55.0	48.4
Open-source Vision-Language Model								
Mulberry-7B	-	63.1	39.6	-	-	-	54.1	-
InternVL2.5-8B	-	64.4	39.5	19.7	34.3	-	-	-
Kimi-VL-16B	-	68.7	44.9	21.4	-	-	-	-
Qwen2-VL-7B	-	58.2	32.5	16.3	30.5	20.2	50.6	34.7
Qwen2.5-VL-7B*	-	68.5	46.8	24.0	<u>42.3</u>	24.4	67.5	45.6
Concurrent Vision-Language Reasoning Models								
R1-VL-7B	270K	63.5	40.0	24.7	7.8	8.3	54.7	33.2
R1-Onevision-7B	165K	64.1	46.4	29.9	21.6	20.8	65.6	41.4
VLAA-Thinker-7B*	150K	<u>70.0</u>	<u>48.6</u>	24.9	42.0	<u>25.5</u>	<u>68.4</u>	<u>46.6</u>
OpenVLThinker-7B*	12K	72.3	50.3	<u>25.9</u>	42.9	26.8	69.1	47.9

framework is based on LLaMA-Factory² for SFT and EasyR1³ for RL. We source our training data from the established LLaVA-OneVision [33] and specifically consider the 14 data sources in overlap with MathV360K [56] (Table 7). We defer the detailed descriptions of evaluation benchmarks, baselines, data and training hyperparameters to Appendix D.

3.1 Main Results

We present our main results in Table 1. As illustrated, OpenVLThinker-7B consistently achieves either the best or second-best scores among open-source LVLMS of comparable scale across all six benchmarks, including concurrent reasoning models. On average, OpenVLThinker attains an accuracy of 46.6%, representing a 2% improvement over the base model and performance comparable to proprietary models such as GPT-4o. Notably, OpenVLThinker exhibits fewer hallucinations and more precise perception than its base model on HallusionBench, improving accuracy by 2.7%. Compared to concurrent reasoning methods that utilize substantially larger datasets for single-iteration SFT and RL, our iterative approach achieves superior results while utilizing only 1/10 of the data scale as used in concurrent works with a single-iteration SFT-RL pipeline.

OpenVLThinker-3B. We additionally train a 3B model using a single iteration of the SFT-RL pipeline, where the training process distills from our 7B model. In Table 2, we compare the performance of our 3B model against current representative models at the same scale, including our base model, Qwen2.5-VL-3B, and the reasoning model VLAA-Thinker-3B, which is trained from the same initial checkpoint as ours. OpenVLThinker-3B achieves the best performance on MathVista and outperforms state-of-the-art 3B reasoning models.

Table 2: Performance of 3B models on MathVista.

Model	Accuracy (%)
R1-VL-2B	52.1
InternVL2.5-4B	60.5
Qwen2.5-VL-3B	62.3
VLAA-Thinker-3B	61.0
OpenVLThinker-3B	63.4

²<https://github.com/hiyouga/LLaMA-Factory>

³<https://github.com/hiyouga/EasyR1>

References

- [1] Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*, 2024.
- [2] AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. claude-3 model card. 2024.
- [3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [4] Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. Mapqa: A dataset for question answering on choropleth maps. *arXiv preprint arXiv:2211.08545*, 2022.
- [5] Hardy Chen, Haoqin Tu, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Vl-thinking: An rl-derived visual instruction tuning dataset for thinkable lvlms. <https://github.com/UCSC-VLAA/VL-Thinking>, 2025.
- [6] 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 rl-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*, 2025.
- [7] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *arXiv preprint arXiv:2105.14517*, 2021.
- [8] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.
- [9] Zhipeng Chen, Yingqian Min, Beichen Zhang, Jie Chen, Jinhao Jiang, Daixuan Cheng, Wayne Xin Zhao, Zheng Liu, Xu Miao, Yang Lu, Lei Fang, Zhongyuan Wang, and Ji-Rong Wen. An empirical study on eliciting and improving rl-like reasoning models, 2025. URL <https://arxiv.org/abs/2503.04548>.
- [10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.
- [11] Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*, 2024.
- [12] Yifan Du, Zikang Liu, Yifan Li, Wayne Xin Zhao, Yuqi Huo, Bingning Wang, Weipeng Chen, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. Virgo: A preliminary exploration on reproducing o1-like mllm. *arXiv preprint arXiv:2501.01904*, 2025.
- [13] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [14] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.
- [15] Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, Kaipeng Zhang, Wenqi Shao, Chao Xu, Conghui He, Junjun He, Hao Shao, Pan Lu, Hongsheng Li, and Yu Qiao. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. In *International Conference on Machine Learning (ICML)*, 2024.

- [16] Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach, 2025. URL <https://arxiv.org/abs/2502.05171>.
- [17] Google. Gemini 2.5 pro, May 2025. URL <https://deepmind.google/technologies/gemini/>.
- [18] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.
- [19] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [20] Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. *arXiv preprint arXiv:2412.05237*, 2024.
- [21] Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*, 2025.
- [22] Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, et al. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. *arXiv preprint arXiv:2504.11456*, 2025.
- [23] Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. V-star: Training verifiers for self-taught reasoners. *arXiv preprint arXiv:2402.06457*, 2024.
- [24] Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, and Heung-Yeung Shum Xiangyu Zhang. Open-reasoner-zero: An open source approach to scaling reinforcement learning on the base model, 2025.
- [25] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv preprint arXiv:2406.09403*, 2024.
- [26] 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, 2025. URL <https://arxiv.org/abs/2503.06749>.
- [27] Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei Qin, Weizhe Yuan, and Pengfei Liu. O1 replication journey – part 2: Surpassing o1-preview through simple distillation, big progress or bitter lesson?, 2024. URL <https://arxiv.org/abs/2411.16489>.
- [28] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [29] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
- [30] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017.

- [31] Seungone Kim, Se Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. The CoT collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12685–12708, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.782. URL <https://aclanthology.org/2023.emnlp-main.782/>.
- [32] Sicong Leng, Jing Wang, Jiayi Li, Hao Zhang, Zhiqiang Hu, Boqiang Zhang, Hang Zhang, Yuming Jiang, Xin Li, Deli Zhao, Fan Wang, Yu Rong, Aixin Sun[†], and Shijian Lu[†]. Mmr1: Advancing the frontiers of multimodal reasoning. <https://github.com/LengSicong/MMR1>, 2025.
- [33] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [34] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.
- [35] Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *ACL*, 2023.
- [36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2023.
- [37] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.
- [38] Runze Liu, Junqi Gao, Jian Zhao, Kaiyan Zhang, Xiu Li, Biqing Qi, Wanli Ouyang, and Bowen Zhou. Can 1b llm surpass 405b llm? rethinking compute-optimal test-time scaling, 2025. URL <https://arxiv.org/abs/2502.06703>.
- [39] Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. Noisyrollout: Reinforcing visual reasoning with data augmentation. *arXiv preprint arXiv:2504.13055*, 2025.
- [40] Zichen Liu, Changyu Chen, Wenjun Li, Tianyu Pang, Chao Du, and Min Lin. There may not be aha moment in rl-zero-like training — a pilot study, 2025. Notion Blog.
- [41] Zichen Liu, Changyu Chen, Wenjun Li, Tianyu Pang, Chao Du, and Min Lin. There may not be aha moment in rl-zero-like training — a pilot study. <https://oatllm.notion.site/oat-zero>, 2025. Notion Blog.
- [42] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.
- [43] 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.
- [44] Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl, 2025. Notion Blog.
- [45] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhui Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

- [46] Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, Wayne Xin Zhao, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems, 2024. URL <https://arxiv.org/abs/2412.09413>.
- [47] Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18798–18806, 2024.
- [48] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling, 2025. URL <https://arxiv.org/abs/2501.19393>.
- [49] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- [50] Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. *Advances in Neural Information Processing Systems*, 37:116617–116637, 2024.
- [51] Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, and Pengfei Liu. O1 replication journey: A strategic progress report – part 1, 2024. URL <https://arxiv.org/abs/2410.18982>.
- [52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.
- [53] Amrith Setlur, Nived Rajaraman, Sergey Levine, and Aviral Kumar. Scaling test-time compute without verification or rl is suboptimal, 2025. URL <https://arxiv.org/abs/2502.12118>.
- [54] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, YK Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- [55] 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.
- [56] Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*, 2024.
- [57] Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, et al. Beyond human data: Scaling self-training for problem-solving with language models. *arXiv preprint arXiv:2312.06585*, 2023.
- [58] Nishad Singhi, Hritik Bansal, Arian Hosseini, Aditya Grover, Kai-Wei Chang, Marcus Rohrbach, and Anna Rohrbach. When to solve, when to verify: Compute-optimal problem solving and generative verification for llm reasoning. *arXiv preprint arXiv:2504.01005*, 2025.
- [59] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.
- [60] OpenThoughts Team. Open Thoughts. <https://open-thoughts.ai>, January 2025.
- [61] Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL <https://qwenlm.github.io/blog/qwq-32b/>.

- [62] Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*, 2025.
- [63] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [64] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [65] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025.
- [66] 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, 2024.
- [67] 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.
- [68] Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, et al. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*, 2025.
- [69] Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*, 2025.
- [70] Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning, 2025.
- [71] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.
- [72] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [73] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization, 2025. URL <https://arxiv.org/abs/2503.10615>.
- [74] Huanjin Yao, Jiaying 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.
- [75] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning, 2025.
- [76] Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in llms, 2025.
- [77] Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

- [78] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models, 2024. URL <https://arxiv.org/abs/2401.10020>.
- [79] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.
- [80] Dan Zhang, Sining Zhou, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. Rest-mcts*: Llm self-training via process reward guided tree search. *Advances in Neural Information Processing Systems*, 37:64735–64772, 2025.
- [81] Di Zhang, Jianbo Wu, Jingdi Lei, Tong Che, Jiatong Li, Tong Xie, Xiaoshui Huang, Shufei Zhang, Marco Pavone, Yuqiang Li, et al. Llama-berry: Pairwise optimization for o1-like olympiad-level mathematical reasoning. *arXiv preprint arXiv:2410.02884*, 2024.
- [82] Hanning Zhang, Jiarui Yao, Chenlu Ye, Wei Xiong, and Tong Zhang. Online-dpo-r1: Unlocking effective reasoning without the ppo overhead, 2025. Notion Blog.
- [83] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025.
- [84] Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*, 2024.
- [85] 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.
- [86] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.
- [87] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibe Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191, 2023.
- [88] Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero’s “aha moment” in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*, 2025.

A Related Work

A.1 Complex Chain-of-Thought Reasoning

Since the introduction of OpenAI’s O1 model [29], researchers have shown strong interest in reproducing and enhancing the complex reasoning capabilities of LLMs [27, 46, 51, 80, 81], partly due to its superior performance on mathematical benchmarks. [19] introduce the open-source *DeepSeek-R1* model and investigate how RL with verifiable rewards can promote advanced chain-of-thought reasoning and reflective behaviors. This development inspired a line of research focused on open-source reproduction [24, 40, 44, 60, 82] and the analysis of such complex reasoning in mathematical problem solving [9, 70, 75, 76]. In parallel, several recent studies have similarly explored the effects of test-time scaling on encouraging more complex model reasoning behaviors [16, 38, 48, 53, 58, 84]. However, the majority of research have significantly advanced text-based reasoning, and development of vision-language reasoning is much more initial.

A.2 Vision-Language Reasoning Model

Recent advancements in large vision-language models (LVLMs) stem from open-source LLMs [13, 63, 64, 72] and text-aligned image encoders [34, 52]. Integrating these components has enabled LVLMs to follow diverse visual instructions and generate meaningful responses [3, 10, 14, 15, 36, 37]. Parallel to the model development, researchers have also been interested in eliciting CoT reasoning chains from LVLMs via prompting [25, 47, 86, 87] or fine-tuning [12, 20, 62, 71]. These reasoning models remain mostly on a shallow level of common step-by-step prompting, without self-reflections or self-verifications.

Concurrent work. Very recently, many studies have started exploring how to equip LVLMs with R1-like reasoning capabilities through distillation from text-only reasoning models [5, 26, 32, 73] or directly rely on RL [42, 88] for self-exploration. Further advancements [6, 39, 45, 55, 65, 68, 69, 83] have focused on improving performance in visual math reasoning, which marks the transition from early-stage exploration to more effective complex vision-language reasoning. Please note that most of these works are within the two months before the submission date, and some of them do not even have associated technical reports available yet. Our work aligns with these studies and contributes unique insights into the role of SFT for complex reasoning, along with an iterative SFT-RL framework to further advance research in this direction.

B Preliminaries

An LLM is defined by a probability distribution p_{θ} , parameterized by model weights θ . Given a prompt sequence $\mathbf{x} = [x_1, \dots, x_n]$, the model generates a response sequence $\mathbf{y} = [y_1, \dots, y_m]$, where x_i and y_j represent individual tokens. The response $\mathbf{y}$ is sampled from the conditional distribution $p_{\theta}(\cdot|\mathbf{x})$, factorized as $p_{\theta}(\mathbf{y}|\mathbf{x}) = \prod_{j=1}^m p_{\theta}(y_j|\mathbf{x}, y_1, \dots, y_{j-1})$.

Supervised Fine-Tuning (SFT). SFT is typically applied to specialize LLMs for a particular task or domain. This process updates the model parameters θ by providing example responses of desired behavior to the input instructions. Concretely, Given a dataset $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^N$, where $\mathbf{x}^{(i)}$ is the prompt sequence and $\mathbf{y}^{(i)}$ is the desired response sequence. We update θ to maximize the likelihood of producing $\mathbf{y}^{(i)}$ given $\mathbf{x}^{(i)}$. Formally, $\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{i=1}^N \log p_{\theta}(\mathbf{y}^{(i)}|\mathbf{x}^{(i)})$. By minimizing the loss, the model learns to produce responses more aligned with the labeled examples.

Reinforcement Learning (RL). RL approaches fine-tune LLMs via human preferences modeled under the Bradley-Terry model [1, 11, 49, 54]: $p(\mathbf{y}_w \succ \mathbf{y}_l | \mathbf{x}) = \sigma(r(\mathbf{x}, \mathbf{y}_w) - r(\mathbf{x}, \mathbf{y}_l))$, where $\mathbf{y}_w$ and $\mathbf{y}_l$ denote preferred and dispreferred responses, respectively, and $\sigma(t) = 1/(1 + e^{-t})$ is the sigmoid function. The common RL objective under the Bradley-Terry assumption of the reward model $r(\mathbf{x}, \mathbf{y})$ is thus

$$\max_{\theta} \left[\mathbb{E}_{\mathbf{x}, \mathbf{y} \sim p_{\theta}} [r(\mathbf{x}, \mathbf{y})] - \beta \mathbb{E}_{\mathbf{x}} [\text{KL}(p_{\theta}(\cdot|\mathbf{x}) \| p_{\text{ref}}(\cdot|\mathbf{x}))] \right],$$

where $\beta > 0$ is the KL penalty coefficient. Under this framework, [54] introduced Group Relative Policy Optimization (GRPO) by sampling a group of response trajectories $\{\mathbf{o}_i\}_{i=1}^G$ from the old

policy model θ_{old} for each query $\mathbf{x}$, with the objective as maximizing:

$$\mathbb{E} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|\mathbf{o}_i|} \sum_{t=1}^{|\mathbf{o}_i|} \min \left(\frac{p_{\theta}(o_{i,t} | \mathbf{x}, \mathbf{o}_{i,<t})}{p_{\theta_{\text{old}}}(o_{i,t} | \mathbf{x}, \mathbf{o}_{i,<t})} \hat{A}_{i,t}, \text{clip} \left(\frac{p_{\theta}(o_{i,t} | \mathbf{x}, \mathbf{o}_{i,<t})}{p_{\theta_{\text{old}}}(o_{i,t} | \mathbf{x}, \mathbf{o}_{i,<t})}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}_{i,t} \right) \right] - \beta \mathbb{D}_{\text{KL}}[p_{\theta} \parallel p_{\theta_{\text{ref}}}], \quad (1)$$

where $\epsilon > 0$ is a hyperparameter bounding the clipping range, $\beta > 0$ balances the KL-penalty term $\mathbb{D}_{\text{KL}}[\pi_{\theta} \parallel \pi_{\text{ref}}]$ against the advantage-weighted policy update, and θ_{old} is the old policy model. Here, the advantage $\hat{A}_{i,t} = \tilde{r}_i = (r_i - \text{mean}(r)) / \text{std}(r)$ is set as the normalized reward at group level.

C Additional Method Details

C.1 Two-Stage Source-Based Curriculum RL

To ensure effective exploration during reinforcement learning (RL), we assess the difficulty of data sources, aiming to provide data that is challenging yet appropriate for the model’s proficiency level. Specifically, we utilize GPT-4o to rate the difficulty of five representative examples drawn from various data sources such as FigureQA [30], MapQA [4], and GeoQA [7], in a similar fashion to the text-based evaluation in DeepMath-103K [22]. Additionally, we employ the base model, Qwen2.5-VL-7B⁴, to obtain its error rates as a complementary difficulty indicator. We standardize independently using z-score normalization for both the GPT-4o rating and base model error rates and compute the average of the two. Based on this composite score, we categorize the data sources into Easy, Medium, and Hard groups via k-means clustering in 1d space. With these categories, we construct two difficulty-specific datasets: $\mathcal{D}_{\text{RL(Medium)}}$ and $\mathcal{D}_{\text{RL(Hard)}}$. Our curriculum training thus proceeds in two stages within one iteration, sequentially training on $\mathcal{D}_{\text{RL(Medium)}}$ and $\mathcal{D}_{\text{RL(Hard)}}$.

D Experiment Settings

Training data. Based on our preliminary experiments, we equally draw 500 examples from each source to form the SFT seed dataset of 7K examples, where for each iteration we collect distillation data via rejection sampling, resulting in a final 3K SFT data. We then classify the data sources into easy, medium and hard (as detailed in Table 7). We construct the 3K medium-level RL training data from the 5 sources that we identified as medium difficulty. Finally, we construct 6K hard-level RL training data from the 3 most difficult sources, summing up to 12K data in total for each iteration that trains from the base model.

Evaluation. Our evaluation employs exact matching and a grader function from MathRuler⁵. We use the same inference hyperparameter as suggested by Qwen and recovered Qwen2.5-VL-7B’s reported results on MathVista at 68.5%. The hyperparameters are detailed in Table 3. We employ six established benchmarks to examine model’s ability thoroughly:

- Math reasoning: MathVista [43], MathVerse [85] and MathVision [66]. The three benchmarks evaluate how LVLs interpret and reason with diagrams in visual math problems through both multiple-choice and free-form questions.
- General reasoning: MMMU-Pro [79] and EMMA [21]. MMMU-Pro spans 30 subjects across 183 subfields, including business, medicine, and science. EMMA evaluates in physics, chemistry, coding, and math.
- Perception: HallusionBench [18], designed to evaluate LVLs’ susceptibility to language hallucination and visual illusion.

Baselines. We evaluate the non-reasoning base model Qwen2.5-VL-7B as a primary baseline to demonstrate the improvements introduced by our method. Additionally, we include the reported performance of proprietary models, including GPT-4o [28] and Claude-3.5-Sonnet [2], alongside open-source LVLs such as Mulberry-7B [74], InternVL2.5-8B [8], Kimi-VL-16B [59], and Qwen2-VL-7B [67], as reference points. Crucially, to highlight the effectiveness of our iterative SFT-RL training strategy, we compare our model with concurrent approaches employing a single round of SFT distillation and RL at the same model scale (7B), yet utilizing significantly larger training

⁴In alignment with previous RL reasoning research [75, 76], we choose the base model from Qwen2.5 family for their strong general capability obtained in pre-training.

⁵<https://github.com/hiyouga/MathRuler>

datasets. These concurrent models include R1-VL-7B [83], R1-Onevision-7B [73], and VLAA-Thinker-Qwen2.5VL-7B [6]. Notably, R1-Onevision and VLAA-Thinker-Qwen2.5VL-7B also start from the same base model (Qwen2.5-VL-7B) as ours, using 165K and 150K total data, respectively. In contrast, our model achieves better performance with only 12K training samples from the base model.

Hyperparameters. We thank LLaMA-Factory⁶ and EasyR1⁷ for open-sourcing the training framework that we used for SFT and GRPO. In Table 4 and 5, we detail the hyperparameters that we used for SFT, GRPO and inference. We further lay out the prompts we used for generating image captions. Experiments were conducted on GPU clusters to the similar level of NVIDIA H100 80GB GPU. SFT/Distillation requires 30 minutes and RL requires 20 hours for each iteration. In addition, distillation data generation with verification requires about 8 hours.

Table 3: Inference hyperparameters.

max_new_tokens	2048
top_p	0.001
top_k	1
temperature	0.01
repetition_penalty	1.0

Table 4: Supervised fine-tuning hyperparameters.

Data type	bf16
Learning rate	5e-7
Global batch size	32
Scheduler	Cosine
Warmup ratio	0.1
Num train epochs	1
Image max pixels	262144

Table 5: GRPO hyperparameters.

Rollout batch size	512
Global batch size	128
Max grad norm	1.0
Data type	bf16
Learning rate	1e-6
Weight decay	1e-2
Warmup ratio	0.0
Rollout temperature	1.0

Prompt for image description generation

Please provide a detailed description of this image that could help another AI model understand it completely. Be specific and comprehensive while maintaining natural language flow.

E Additional Experiments

Distillation at iteration 1. At SFT-Iter1, we utilized the base model Qwen2.5-VL-7B to generate image descriptions and obtained R1-like reasoning from QwQ-32B through rejection sampling. A common problem for distillation observed in text-only math reasoning is the overly long reasoning length coupled with unnecessary repetitions of reflections [41, 76]. We observed similarly that these initial reasoning traces were often excessively verbose, partly due to information loss during

⁶<https://github.com/hiyouga/LLaMA-Factory>

⁷<https://github.com/hiyouga/EasyR1>

Table 6: Performance on the MathVista benchmark comparing different SFT data-filtering strategies. Removing the most repetitive keywords in data can mitigate repetitive reflections after SFT.

Model Variant	Accuracy (%)
Qwen2.5-VL-7B	68.5
Vanilla	57.5
Filtered	58.7
Truncated (SFT-Iter1)	63.4

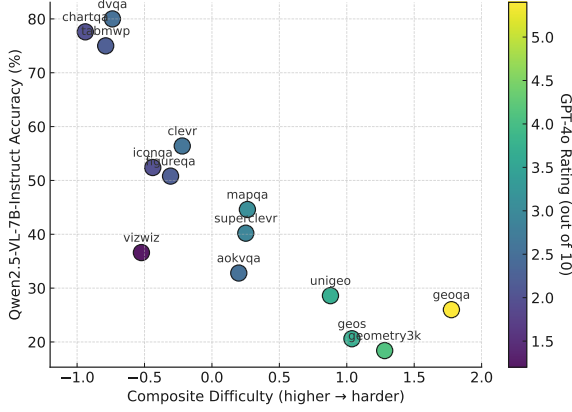


Figure 4: Data source difficulty based on base model accuracy and GPT-4o rating.

Table 7: Categorization of data sources by composite difficulty score using k-means with $k=3$. The geometry question sources all fall into the hard category.

Easy	Medium	Hard
ChartQA	FigureQA	UniGeo
IconQA	CLEVR	GEOS
VizWiz	A-OKVQA	Geometry3K
TabMWP	SuperCLEVR	GeoQA
DVQA	MapQA	

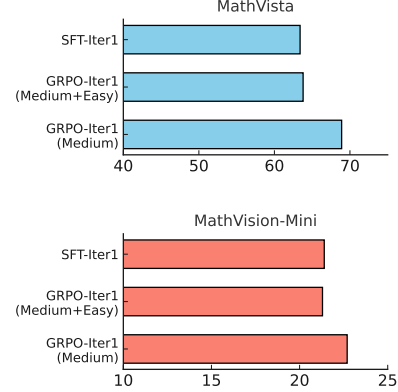


Figure 5: Performance at GRPO-Iter1 using data from different difficulty sources, at the same scale of 3K.

image-to-caption conversion. Consequently, post-SFT reasoning became increasingly repetitive with unproductive self-reflections (see Appendix F for an illustration). To address this, we evaluated two filtering strategies: (1) discarding samples with reasoning traces exceeding 500 words, and (2) truncating reflections by splitting traces of at specific keywords that were overly repetitive in data (“Wait,” “But wait,” and “But the question”) and discarding subsequent segments while preserving the final answer. The latter approach was ultimately adopted to prevent the model from internalizing reflection loops, while preserving the reasoning action at a reasonable frequency. Table 6 compares models trained on original versus processed data.

Data source difficulty. We conducted a quantitative analysis to categorize the data sources based on difficulty. Applying k-means clustering (with $k = 3$) to our composite difficulty score as described in Section 2 allowed us to clearly identify three distinct difficulty pools, as shown in Table 7. We visualize the difficulty scores for each source in Figure 4. In Figure 5, we show the performance of GRPO-Iter1 when drawing 3K data from either (1) 10 data sources classified as either Easy or Medium, or (2) 5 data sources classified as Medium. We observe that RL training with easy-level data results in ineffective performance gain as compared to sourcing from medium-level data only. This finding aligns with concurrent algorithmic efforts such as DAPO [77] in the text-only domain for improving GRPO by dynamically filtering out overly-easy examples.

Curriculum RL to maximize utilization of challenging data. Figure 6 investigates the impact of incorporating challenging training data (e.g., geometry datasets) at iteration 1. On the left panel, we illustrate the absolute performance gains transitioning from SFT-Iter1 to GRPO-Iter1 (medium difficulty), and subsequently from GRPO-Iter1 (medium) to GRPO-Iter1 (hard). Training on these harder datasets yields substantial improvements on more difficult benchmarks, such as MathVision, while not significantly affecting performance on easier benchmarks like MathVista. On the right panel, we further compare our two-stage, source-based curriculum RL approach against training solely on hard data. The results indicate that initiating RL with moderately challenging (medium difficulty) data and subsequently progressing to harder datasets provides optimal performance improvements.

Iterative progression. Building upon the performance improvements shown in Figure 3, we further analyze changes in reasoning length across iterations, as illustrated in Figure 7. Our results indicate

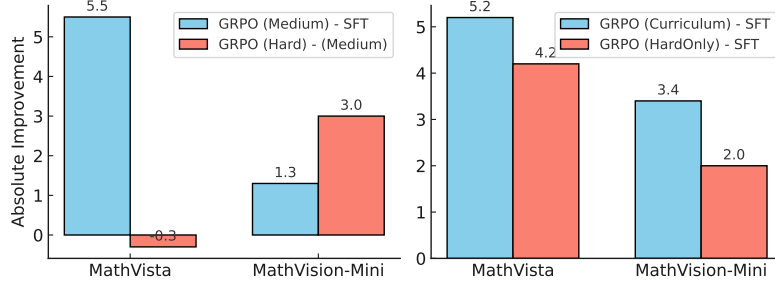


Figure 6: Absolute performance gain achieved at iteration 1. The round 2 RL training on hard data provides more significant performance gain on harder benchmarks such as MathVision. Moreover, if RL training with the hard data only yield less improvement than our two-stage curriculum RL.

that the reasoning model consistently utilizes more words at inference time compared to the base non-reasoning model, without becoming excessively repetitive. Notably, the largest increase in reasoning length occurs at Iteration 1, with subsequent iterations gradually adopting more concise reasoning. This progression suggests an increasingly efficient utilization of reflective reasoning, engaging reflections primarily when beneficial. In Appendix F (Figure 8 and 9), we show reasoning examples that our SFT-ed model was incorrect while our RL-ed model was correct.

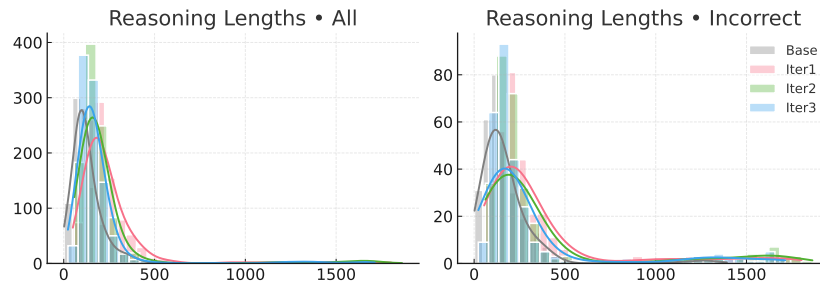


Figure 7: Distribution of reasoning length (number of words) across iterations of training. While our trained reasoning model across iterations all tend to reason longer than the base model, iterative training resulted in gradually more concise length, possibly due to reduced repetitive reflections.

F Output Examples

Repetitive reasoning example after SFT without filtering

But the question is asking for the route and endpoint. So, the route would be the highway number, and the endpoint would be the destination.

Wait, but the signs are arranged vertically. So, the top signs are the routes, and the bottom ones are the destinations. So, if you're at the junction and you're facing the signs, you can see both. So, the route is the highway number, and the endpoint is the destination.

But the question is phrased as "the route and endpoint reached by continuing in a straight line." So, if you're at the junction and you're facing the signs, you can see both the routes and the destinations. So, the route would be the highway number, and the endpoint would be the destination.

Wait, but the instruction is a bit ambiguous. It says "continuing in a straight line," which might mean following the route indicated by the sign. So, if you're at the junction and you're facing the signs, you can see both the routes and the destinations. So, the route is the highway number, and the endpoint is the destination.

But the question is asking for the route and endpoint. So, the route would be the highway number, and the endpoint would be the destination.

Wait, but the signs are arranged vertically. So, the top signs are the routes, and the bottom ones are the destinations. So, if you're at the junction and you're facing the signs, you can see both. So, the route is the highway number, and the endpoint is the destination.

But the question is phrased as "the route and endpoint reached by continuing in a straight line." So, if you're at the junction and you're facing the signs, you can see both the routes and the destinations. So, the route would be the highway number, and the endpoint would be the destination.

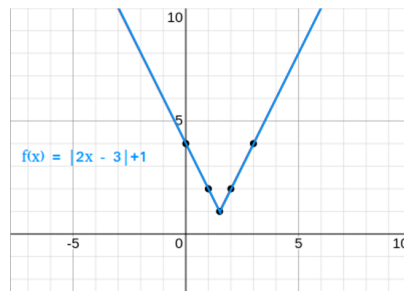


Figure 8: The image to the visual reasoning task that we show for SFT-Iter3 and GRPO-Iter3 output.

Example of SFT-Iter3 incorrect output and GRPO-Iter3 Correct output to the same question

Question: The derivative of $f(x)$ at $x=2$ is ____ that at $x=5$
 Choices: (A) larger than (B) equal to (C) smaller than

SFT-Iter3

<think>

Okay, let me see. The question is asking about the derivative of the function $f(x) = |2x - 3| + 1$ at $x=2$ compared to $x=5$. First, I need to recall what the derivative represents. The derivative gives the slope of the tangent line at a point on the function. So I need to look at the slopes at those two points.

</think>

The function ($f(x) = |2x - 3| + 1$) is an absolute value function, which means it has a V-shape. The vertex of this V-shape occurs where the expression inside the absolute value is zero, i.e., ($2x - 3 = 0$), which gives ($x = \frac{3}{2}$). The derivative of ($f(x)$) will be zero at the vertex because the function changes direction there. For ($x < \frac{3}{2}$), the function is decreasing, and for ($x > \frac{3}{2}$), it is increasing. The slopes on either side of the vertex are negative and positive, respectively. At ($x = 2$) (which is to the right of the vertex), the function is increasing, so the derivative is positive. At ($x = 5$) (which is further to the right), the function is also increasing, so the derivative is positive but steeper because the function is increasing faster. Since the function is increasing faster at ($x = 5$) than at ($x = 2$), the derivative at ($x = 5$) is larger.

<answer>(C)</answer>

=====

GRPO-Iter3

<think>

Okay, let's see. The question is asking about the derivative of the function $f(x) = |2x - 3| + 1$ at $x=2$ compared to $x=5$. First, I need to recall what the derivative represents. The derivative at a point gives the slope of the tangent line at that point. So I need to look at the slopes at $x=2$ and $x=5$.

</think>

The function ($f(x) = |2x - 3| + 1$) is an absolute value function shifted up by 1. The absolute value function ($|2x - 3|$) has a V-shape with a vertex at ($x = \frac{3}{2}$). The derivative of ($|2x - 3|$) is (2) for ($x > \frac{3}{2}$) and (-2) for ($x < \frac{3}{2}$). Adding 1 to the function doesn't affect the derivative's magnitude but shifts the y-values up. At ($x = 2$), which is greater than ($\frac{3}{2}$), the derivative is (2). At ($x = 5$), which is also greater than ($\frac{3}{2}$), the derivative is also (2). Since the derivative is the same at both points, the slopes are equal.

<answer>(B)</answer>

Ground truth: (B)

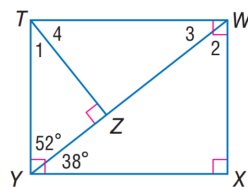


Figure 9: The image to the task that we show for SFT-Iter3 and GRPO-Iter3 output below.

Example of SFT-Iter3 incorrect output and GRPO-Iter3 Correct output to the same question

Question: Find $m\angle 3$.

Choices: (A) 28 (B) 38 (C) 52 (D) 62

SFT-Iter3

<think>

Okay, let me see. The problem is asking for the measure of angle 3 in the given quadrilateral. The hint suggests looking at the angles provided, which are 52 degrees and 38 degrees. Hmm, the quadrilateral is a rectangle, so all angles should be 90 degrees. But the hint is confusing me. Let me check the options again. The angles 52 and 38 are given, but they're not the ones I need to find. The question is about angle 3, which is opposite angle 1. Since opposite angles in a rectangle are equal, angle 3 should be the same as angle 1. But angle 1 is not given directly. Let me see, the hint is saying to look at the angles provided. The 52 and 38 are supplementary angles, which means they add up to 180 degrees. But angle 3 is not supplementary to 52 or 38. It's the opposite angle, so it should be the same as angle 1. Since angle 1 is not given, I need to use the options. The only angle that could be the same as angle 1 is the one that's not given directly. The options are A, B, C, D. The only one that's not supplementary to 52 or 38 is 62. So angle 3 must be 62 degrees.

</think>

The problem states that the quadrilateral is a rectangle, and all angles in a rectangle are 90 degrees. The hint provided the angles 52° and 38° , which are supplementary angles ($52^\circ + 38^\circ = 90^\circ$). However, angle 3 is the opposite angle to angle 1, which means they are equal because opposite angles in a rectangle are congruent. Since angle 1 is not given directly, we need to look at the options provided. The only angle that fits the properties of opposite angles in a rectangle is the one that is not supplementary to the given angles. The options are A (28), B (38), C (52), and D (62). Since angle 1 is not given, the only angle that could be the same as angle 3 is the one that is not supplementary to 52° or 38° . The only angle that fits this criterion is 62° , as it is the only one not listed as supplementary to the given angles.

<answer>(D)</answer>

=====

GRPO-Iter3

<think>

Okay, let's see. The question is asking for the measure of angle 3. The image provided is a rectangle with some angles labeled. The hint suggests looking at the angles given, which are 52 degrees and 38 degrees. Hmm, the rectangle has right angles, so each angle should be 90 degrees. The angles 52 and 38 are supplementary because they add up to 90 degrees ($90 - 52 = 38$). But angle 3 is not one of those. It's the angle opposite the 38-degree angle, so it should be the same. Wait, the question is about angle 3, which is the one opposite the 38-degree angle. So it should be 38 degrees too, right?

</think>

The problem states that angle 3 is opposite the 38-degree angle. In a rectangle, opposite angles are equal. Since angle 3 is opposite the 38-degree angle, angle 3 must also be 38 degrees.

<answer>(B)</answer>

Ground truth: (B)

G Impact Statement

Our work introduces *OpenVLThinker*, an open-source large vision-language model (LVLM) designed explicitly for complex reasoning tasks through a novel iterative self-improvement process integrating supervised fine-tuning (SFT) and reinforcement learning (RL).

Positive societal impacts include significant improvements in multimodal reasoning, beneficial for applications in education, assistive technologies, accessibility, and content moderation, potentially enhancing decision-making processes involving visual and textual data.

However, improved generative capabilities inherently pose potential negative societal impacts, such as misuse for creating deceptive or manipulative content or misinformation. There are also fairness and privacy considerations, as biases in training data or misinterpretation of visual information could unfairly affect specific groups.

To mitigate these risks, responsible deployment practices such as gated model releases, comprehensive documentation, usage guidelines, and continuous monitoring for misuse and biases are essential.