Vision-G1: Towards General Reasoning Vision-Language Models via Reinforcement Learning

Yuheng Zha[1*], Kun Zhou[1*†], Yujia Wu[1], Yushu Wang[1], Jie Feng[1], Zhi Xu[1], Shibo Hao[1], Zhengzhong Liu[3], Eric P. Xing[2,3], Zhiting Hu[1]

[1]UC San Diego, 9500 Gilman Dr, La Jolla, CA 92093 USA
[2]Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213 USA
[3]MBZUAI IFM, 150 Mathilda Pl, Sunnyvale, CA 94086 USA
kuzhou@ucsd.edu

Abstract

Recent vision-language models (VLMs) show strong reasoning capabilities through training with reinforcement learning from verifiable rewards (RLVR). Despite their impressive capabilities, current VLMs focus on a limited range of reasoning tasks, such as mathematical and logical reasoning, due to the lack of readily available verifiable reward data in broader domains. As a result, these models struggle to generalize their reasoning abilities to the wide variety of challenges encountered in real-world environments. To address this limitation, we collect and assemble a comprehensive RL-ready visual reasoning training dataset encompassing 46 datasets across 13 dimensions of 5 domains, covering a wide range of realistic scenarios such as infographic reasoning, mathematical reasoning, spatial reasoning, and general science reasoning. Based on this dataset, we propose an influence function-based data filtering strategy and a multi-round data curriculum method to iteratively strengthen general visual reasoning abilities. Using this approach, we train a general reasoning VLM, namely Vision-G1. Our 7B model achieves state-of-the-art performance across nine visual reasoning benchmarks, surpassing similar-sized VLMs and even GPT-4o and Gemini-1.5 Flash.

Code — https://github.com/yuh-zha/Vision-G1

Introduction

Large language models (LLMs) trained with reinforcement learning (RL) from verifiable rewards, such as DeepSeek R1 (Guo et al. 2025), show strong reasoning capabilities on diverse tasks such as math (Cobbe et al. 2021; Hendrycks et al. 2021) and coding (Jimenez et al. 2024). Following this paradigm, the open source community has proposed additional reasoning LLM training methods (Yu et al. 2025; Hu et al. 2025; Yuan et al. 2025) to advance these capabilities further. It is promising to apply similar methods from pure language models to vision language models (VLMs), enabling VLMs to exhibit strong reasoning capabilities on a wide range of visual reasoning tasks. While the common practice for training vision-language models (Bai et al. 2025; Chen et al. 2024d; Liu et al. 2023a) involves only supervised fine-tuning after pre-training, there have already been some initial attempts to post-train VLMs with reinforcement learning from verifiable rewards to enhance their visual perception (Liu et al. 2025; Shen et al. 2025) and reasoning (Peng et al. 2025; Meng et al. 2025a; Du et al. 2025; Yang et al. 2025; Zhan et al. 2025; Huang et al. 2025; Wang et al. 2025b,a; Team et al. 2025) capabilities. For example, by collecting K12-level exam questions with verifiable answers, MM-Eureka (Meng et al. 2025a) trains a VLM to improve its math and science-related reasoning capabilities.

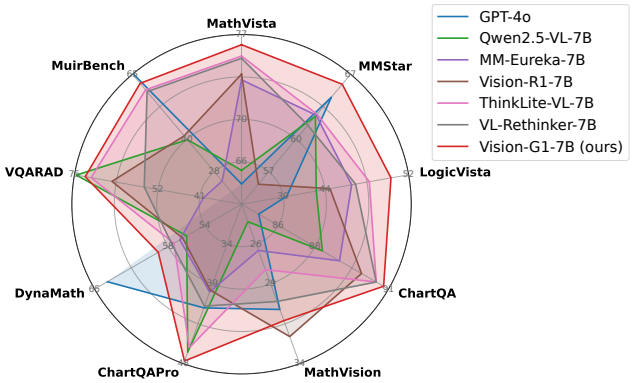

Figure 1: Performance comparison between Vision-G1 (ours) and other methods on ten general visual reasoning datasets.

Despite their success, VLMs still struggle with general visual reasoning tasks, which usually require multiple reasoning steps involving logical, commonsense, and physical knowledge (Zhang et al. 2024b; Xu et al. 2025; Chen et al. 2024a). Unlike humans, who gain problem-solving abilities by applying prior knowledge and interacting extensively with diverse real-world environments, VLMs have limited access to such experiences. In the current reinforcement training paradigm, VLMs mainly interact with restricted data types and lack real-world interactions, making it difficult for them to gain general visual reasoning capabilities.

In this work, we train a general reasoning vision language model with reinforcement learning from verifiable rewards. To generalize the reasoning capability to broader domains,

*These authors contributed equally.
†Corresponding Author

we build a large RL-ready training dataset covering 5 domains: infographic reasoning, mathematical reasoning, cross-image reasoning, spatial reasoning, and knowledge-specific reasoning. It consists of 46 visual reasoning datasets and 13 sub-tasks in total. For each data source, we filter out instances with non-verifiable answers (*e.g.,* open-ended questions), retaining only those with numeric values, multiple-choice options, yes/no answers, or other single-word ground truths.

However, the collected raw datasets generally contain instances of varying quality and quantity. Simply mixing them admits low-quality instances (*e.g.,* too easy or too hard), impeding effective learning of general visual reasoning knowledge. Meanwhile, existing approaches (Shen et al. 2025; Peng et al. 2025) largely rely on heuristic strategies or human-crafted features for filtering, requiring specific manual designs and limiting adaptability to the above heterogeneous datasets. To address this, we propose a general data filtering method based on influence function (Xia et al. 2024; Pruthi et al. 2020), designed to remove unhelpful instances from the RL training data of general reasoning VLMs. Concretely, we consider both the influence on training and target tasks, and use the estimation function to filter negative instances.

To fully learn the reasoning capability from the filtered multi-source datasets, we further design a data curriculum strategy for multi-round RL training. For the data curriculum, we propose a data selection strategy that estimates instance difficulty using the average rollout accuracy per question from the VLM in the previous training round. Moderately difficult instances are selected to match the current VLM's capability. We iteratively apply the data selection and RL training to gradually improve our Vision-G1's general reasoning capability. Our Vision-G1-7B achieves state-of-the-art performance among competitive baselines on 17 benchmarks, spanning comprehensive visual reasoning, math-related reasoning, and domain-specific reasoning tasks.

Related Work

Vision-Language Models. Large language models continue to advance, with GPT-4 (OpenAI 2024a) and Qwen-2.5-VL (Bai et al. 2025) exhibiting emergent skills such as in-context learning and sophisticated reasoning. Recent work has further strengthened both perception and reasoning. LLaVA-NeXT, for example, supports variable-resolution input by tiling images into adaptive grids (Liu et al. 2023a), whereas Qwen2-VL introduces M-RoPE, a refined rotary position encoding that unifies spatial and temporal cues for images and video (Bai et al. 2025). Several systems treat pictures and clips within a single architecture and merge their instruction data during fine-tuning (Xu et al. 2024; Zhu et al. 2025). On the reasoning front, models such as QvQ (Qwen Team 2024) and Virgo (Du et al. 2025) push performance on complex tasks by generating long reasoning chains.

Reasoning Vision-Language Models. Building on breakthroughs in large reasoning language models such as OpenAI o1 (OpenAI 2024b) and DeepSeek-R1 (Guo et al. 2025), recent work has turned to strengthening the reasoning capabilities of Vision-Language Models (VLMs). Early approaches (Xu et al. 2024; Du et al. 2025) assemble multi-modal Chain-of-Thought (CoT) datasets and employ supervised fine-tuning to boost the reasoning ability of VLMs. Motivated by the success of reinforcement learning techniques (Shao et al. 2024; DeepSeek-AI et al. 2025), recent studies have used RL with task-specific, verifiable reward schemes (*e.g.,* answer accuracy and detection IoU) to provide supervisory signals (Chen et al. 2025b; Liu et al. 2025), improving VLM reasoning and exhibiting remarkable performance. However, existing work has found that relying solely on RL often fails to elicit the long chain-of-thought reasoning ability of VLMs. To address this, supervised fine-tuning (Meng et al. 2025a; Chen et al. 2025a) and special prompting mechanisms (Wei et al. 2023; Zhang et al. 2024b) are proposed to encourage the long CoT generation style.

Data Selection for Language Model Training. To train large language models (LLMs) and vision-language models (VLMs), choosing the right data is always critical. Existing data selection methods (Zhou et al. 2024; Xia et al. 2024) focus on removing the redundant or harmful instances to reduce the training cost and improve the stability. Early work (Zhou et al. 2024; Chen et al. 2023) mostly relies on human experience to design heuristic rules, and shows that a high-quality training small dataset is able to learn specific capabilities, *e.g.,* instruction following and human alignment. Subsequent methods leverage the features that can be computed by simple metrics or LLMs (*e.g.,* length, complexity, and diversity), for data value estimation and selection (Jain et al. 2023; Liu et al. 2023c; Zhuo et al. 2024; Muennighoff et al. 2023). However, the above features need specific designs for diverse tasks, making them hard to handle a highly heterogeneous mix of multi-task datasets. To solve it, influence function methods (Pruthi et al. 2020) have been proposed, which can estimate the influence of each training instance on other ones. Recent work has simplified the influence estimation function into a simple gradient similarity computation formula, and exhibited remarkable performance on text and visual instruction selection (Xia et al. 2024; Zhou et al. 2024). In this work, we utilize the influence function for filtering low-quality instances. Besides, we also use the rollout accuracy to estimate the difficulty for training data selection in the multi-round RL training process.

Preliminary

Vision-Language Models. VLM (Li et al. 2025; Liu et al. 2023b) typically consists of a pre-trained visual encoder to process image or video input into visual tokens, and an LLM to read the text and visual tokens for generating the text output. It has undergone large-scale supervised fine-tuning on massive image caption and visual instruction data, which can follow visual instructions and solve simple visual reasoning questions. But for more complex visual reasoning tasks, VLMs may still suffer from generating accurate reasoning steps to reach the answer (Zhang et al. 2024b; Xu et al. 2024).

Task Definition. Given a VLM, we aim to perform RL training on it to improve its general visual reasoning ability, *e.g.,* visual scene reasoning (Antol et al. 2015; Hudson and Manning 2019), multimodal math problems solving (Lu et al.

2024; Wang et al. 2024b), and multi-image relation reasoning (Zhao et al. 2024; Kazemi et al. 2024). To achieve it, it is necessary to construct a RL-ready comprehensive visual reasoning dataset, and manage the learning process to fully use it for training the VLM. Therefore, in this paper, we first collect and preprocess a variety of datasets from different tasks and domains, then devise a general influence function-based method to filter low-quality instances. Next, we devise the multi-round RL training method, which iterates the difficulty-based data selection and RL training, to gradually learn the general knowledge from the dataset.

Approach

In our approach, we first collect multi-domain multi-task datasets to compose the training corpus. Then, we perform multi-round RL to train our Vision-G1, with the influence function based filtering and difficulty-based data selection strategies. We show the details in Figure 2.

Training Corpus Construction

To build a comprehensive general RL training dataset, we collect a mixture of visual reasoning datasets from a variety of domains and tasks, and preprocess the contained instances into a unified format with specific category labels.

Data Collection. To ensure the coverage of visual reasoning knowledge, we collect five hybrid datasets, then add specific reasoning datasets from other domains and tasks.

• *Hybrid Datasets.* We collect four hybrids used for RL in previous work: MM-R1 (Peng et al. 2025), VerMulti (Peng et al. 2025), ThinkLite (Wang et al. 2025b), ViRL39K (Wang et al. 2025a), and MMK12 (Meng et al. 2025a). Each mixes visual-reasoning datasets with different focuses. ViRL39K and MM-R1 cover broad STEM, VerMulti adds geometry/diagram problems, and MMK12 draws from K-12 textbooks.

• *Specific Reasoning Datasets.* To improve the coverage of real-world scenarios, we further add six specific visual reasoning datasets from other domains. Specifically, we select two chart understanding and reasoning datasets (*i.e.,* ChartBench (Xu et al. 2023) and UniChart (Masry et al. 2023)), and three medical domain knowledge reasoning datasets (*i.e.,* SLAKE (Liu et al. 2021), Path-VQA (He et al. 2020), and VQA-RAD (Lau et al. 2018)). Besides, we also add a comprehensive multi-image training dataset (*i.e.,* Mantis-Instruct (Jiang et al. 2024)), which consists of 14 subsets corresponding to different multi-image skills, *e.g.,* co-reference and comparing understanding.

Data Preprocess. To ease the utilization in the RL training process, we preprocess all the data to unify the format and tag the corresponding category and fine-grained dimensions.

• *Format Unification.* For the format, we mainly unify the input prompt and the answer checking rules. Concretely, we use the prompt from ThinkLite (Wang et al. 2025b), which clearly describes the thinking-then-reasoning process. In the prompt, we also require the VLM to generate the answer within a special symbol, to help with answer extraction. For answer checking, we manually craft the rules to guarantee that the accuracy of all the examples can be automatically verified, to support RL training.

• *Category Classification.* To better monitor data proportion and knowledge distribution, we define a category taxonomy and classify all instances from the collected datasets. We first identify five key visual reasoning abilities, namely infographic, mathematical, spatial, knowledge, and cross-image reasoning. Then, we divide 13 fine-grained dimensions, and obtain a hierarchical category taxonomy (Fig. 3) that can connect well with available datasets. Finally, we classify all instances into the above dimensions, by feeding the question with a guidance prompt into Qwen2.5-32B-Instruct.

Multi-round Reinforcement Learning with Data Curriculum

Given the above dataset, we first filter the low-quality data and then perform multi-round RL with data curriculum, for stably boosting the general reasoning ability.

Influence Function based Data Filtering. According to the influence function theory (Pruthi et al. 2020), the influence of an instance z on another one z' for a model parameterized by θ, can be estimated by computing the similarity between their gradients, denoted as:

$$\mathtt{I}(z, z') \propto \mathit{Sim}(\nabla l(z, \theta), \nabla l(z', \theta)). \tag{1}$$

Using it, we can estimate the influence of each training instance by computing its gradient similarity with other instances, to filter the ones with negative influence. Concretely, we devise the RL influence estimation function by considering the influence on all training and target tasks:

$$\mathtt{I}(z) = \frac{1}{|\mathcal{D}_{Tr}|} \sum_{z' \in \mathcal{D}_{Tr}} \mathtt{I}(z, z') + \frac{1}{|\mathcal{D}_{Ta}|} \sum_{z' \in \mathcal{D}_{Ta}} \mathtt{I}(z, z'), \tag{2}$$

where $\mathcal{D}_{Tr}$ and $\mathcal{D}_{Ta}$ denote the instances from all training and target tasks. To prevent the large cost of computing all parameter gradients and RL training, we use a frozen VLM with LoRA (Hu et al. 2022) after fine-tuning on a subset of training data, as the reference model. Then, for each instance, we perform a rollout to generate the solution and compute the gradients of predicting it on LoRA parameters. Next, we perform random projection to obtain the low-dimensional features (Xia et al. 2024), and utilize cosine similarity in Eq. 1 to estimate the influence of each instance. Finally, we filter the low-ranked ones and ensure that the remaining instances of each dimension are uniformly distributed.

Difficulty-based Data Selection. During RL training, too simple or too difficult instances will lead to a large proportion of correct or wrong outputs in rollout results, which can not produce stable rewards for RL training (Wang et al. 2025a). Therefore, we propose a difficulty-based data selection method to ensure a moderate difficulty level in the training data. Concretely, for each instance, we utilize the VLM to perform the rollout k times and compute the average accuracy. We set the threshold that instances with $> 80\%$ and $< 20\%$ average accuracy are too easy and too difficult, respectively. Thus, we select the other ones as training data.

Multi-round RL Training. In the multi-round training process, we iterate the above difficulty-based data selection

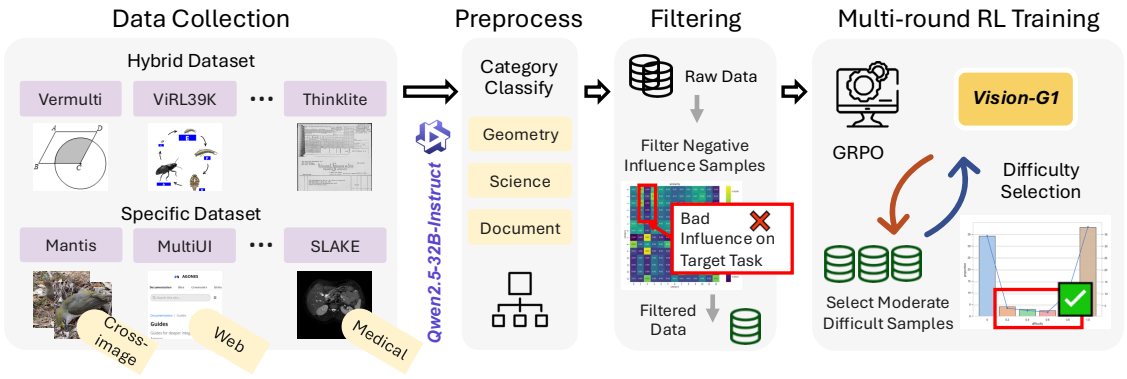

Figure 2: The overview of our approach, consisting of collecting and preprocessing a mixture of heterogeneous datasets, low-quality instances filtering based on influence function, and multi-round RL training with difficulty-based data selection.

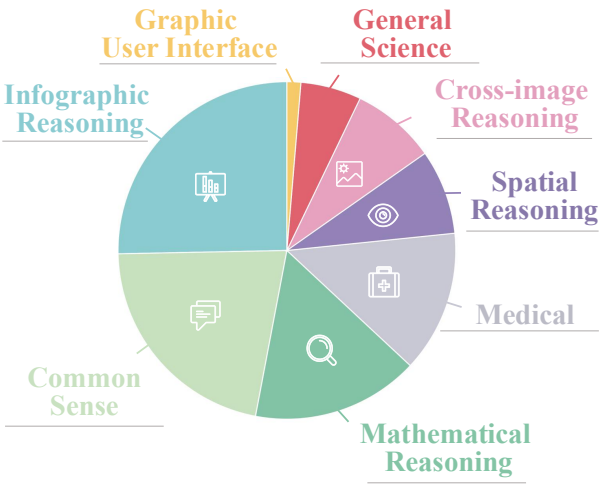

Figure 3: Source dataset distribution of our Vision-G1. KR, IR, MR, CIR, SR denote knowledge, infographic, mathematical, cross-image, and spatial reasoning, respectively.

and VLM RL training until convergence. In each round, we utilize our well-trained VLM in the last round to estimate the difficulty of the untrained data, and then select the samples with a moderate difficulty level for training. Based on the category taxonomy in Fig. 3, we guarantee that the proportion of the different dimension data in the selected training set should be uniform. For RL training, we adopt the Group Relative Policy Optimization (GRPO) algorithm (Shao et al. 2024), and stop training once the results converge.

Experiments

Experimental Setup

In this part, we present the training and evaluation details of our proposed Vision-G1.

Implementation Details. Following the dataset construction pipeline in Section , we create a comprehensive and high-quality RL-ready training dataset with verifiable reward to train our Vision-G1. In the training dataset, math-related problems occupy the half, and the rest of the domains such as chart and medical problems are uniformly distributed. After filtering the low-quality data using the influence function, we have 40k examples in the training data. For RL training, we use an efficient framework `verl`[1] to implement the GRPO algorithm. We train our model on Qwen2.5-VL-7B-Instruct (Bai et al. 2025) for two rounds, with a batch size of 128. For each question in a batch, we randomly sample 32 responses from the model as the rollout results, and use the answer accuracy as the reward of each response. The model is trained with $8\times$ NVIDIA H200 GPUs for around 18 hours. For answer accuracy computation, we use both an open source tool `math-verify`[2] and string matching to compare the ground truth with the model predicted answer. The reward score range is $[0.0, 1.0]$. During evaluation, we use greedy decoding to generate only one response for each question in the benchmark. We use the same answer matching rule to measure the accuracy, and report the model performance using the Pass@1 metric unless otherwise specified.

Evaluation Benchmarks. We evaluate all the models, on a set of comprehensive visual reasoning benchmarks, namely MathVista (Lu et al. 2024), MMMU-Val (Yue et al. 2024a), MMMU-Pro(Yue et al. 2024b), and MMStar (Chen et al. 2024b). The four benchmarks comprehensively evaluate the visual reasoning abilities of VLMs from multiple dimensions, covering visual puzzles, college problems, and science questions. Besides, we also select LogicVista (Xiao et al. 2024) and ChartQA (Masry et al. 2022) due to their comprehensiveness, which test the performance on a variety of logical reasoning and chart or plot understanding tasks, respectively. For mathematical visual reasoning, we include 6 widely-used benchmarks: MathVision (Wang et al. 2024b), MathVerse (Zhang et al. 2024a), OlympiadBench (He et al. 2024), We-Math (Qiao et al. 2024), and DynaMath (Zou et al. 2024). All the above datasets contain the math problems requiring to understand the paired images for solving, spanning from

[1]https://github.com/volcengine/verl
[2]https://github.com/huggingface/Math-Verify

Models	MathVista	MMMU-Val	MMMU-Pro	MMStar	LogicVista	ChartQA	Avg.
GPT-4o	63.8	69.1	51.9	64.7	39.6	85.7	62.5
Claude-3.5	67.7	68.3	51.5	65.1	44.4	90.8	64.6
Gemini-1.5 Flash	58.4	56.1	-	-	40.0	-	-
Gemini-1.5 Pro	63.9	65.8	46.9	59.1	41.8	-	-
Qwen2.5-VL-72B	74.2	68.2	46.2	70.8	-	-	-
InternVL2.5-78B	72.3	70.0	48.6	69.5	-	88.3	-
InternVL3-78B	79.6	72.2	-	72.5	-	89.7	-
VL-Rethinker-32B	78.8	65.6	50.6	-	-	-	-
VL-Rethinker-72B	80.4	68.8	55.9	-	-	-	-
Qwen2-VL-7B	58.2	54.1	30.5	-	-	-	-
Qwen2.5-VL-7B	65.0	**58.6**	38.3	62.8	42.6	88.3	59.3
InternVL2-8B	58.3	51.2	29.0	-	-	-	-
InternVL2.5-8B	64.4	56.0	34.3	-	-	-	-
Llava-OV-7B	63.2	48.8	24.1	-	-	-	-
MiniCPM-V2.6	-	49.8	27.2	60.4	-	-	-
LLaVA-1.6	-	51.1	-	60.7	33.7	-	-
MM-Eureka-7B (Qwen)	73.0	52.7	36.3	62.9	46.2	89.0	60.0
MM-Eureka-8B (Intern)	67.1	49.2	27.8	-	-	-	-
Vision-R1-7B	73.5	49.4	35.2	56.0	44.0	89.9	58.0
R1-VL-7B	63.5	44.5	7.8	-	-	-	-
OpenVLThinker-7B	70.2	52.5	37.3	-	-	-	-
ThinkLite-VL-7B	<u>75.1</u>	53.6	40.1	<u>63.0</u>	<u>48.0</u>	<u>90.5</u>	61.7
VL-Rethinker-7B	74.9	<u>56.7</u>	**41.7**	61.9	46.6	<u>90.5</u>	<u>62.1</u>
Vision-G1 (ours)	**76.1**	53.4	<u>41.2</u>	**66.0**	**50.2**	**90.8**	**63.0**

Table 1: Benchmarking results for general reasoning tasks, where most of the baseline results are sourced from this paper (Zhu et al. 2025) and others from official papers or leaderboards. The best and second-best ones among VLMs with similar scales are marked in bold and underlined, respectively.

simple counting and perception reasoning to complex geometry and combinatorics reasoning tasks. To specifically measure other specific reasoning capabilities, we select the ChartXiv (Wang et al. 2024c) and ChartQAPro (Masry et al. 2025) benchmarks for the chart and plot reasoning ability test, VQA-RAD (Lau et al. 2018), PathVQA (He et al. 2020), and SLAKE (Liu et al. 2021) for measuring the reasoning ability in the medical domain. For multi-image reasoning, we choose MuirBench (Wang et al. 2024a) that contains 12 multi-image understanding tasks.

Baseline Methods. To comprehensively verify the effectiveness of our method, we mainly compare it against VLMs with a similar parameter scale. Specifically, we first select four VLMs with around 7B size, including Qwen2.5-VL-7B (Bai et al. 2025), MiniCPM-V2.6 (Yao et al. 2024) and LLaVA-1.6 (Liu et al. 2024). Among all the above models, Qwen2.5-VL-7B generally performs the best, and has been widely used in existing reasoning VLM work as the backbone. In addition, we also considered the following set of recently proposed reasoning-oriented VLMs that have incorporated Reinforcement Learning (RL) during training, *i.e.,* MM-Eureka-7B (Meng et al. 2025b), MM-Eureka-8B (Meng, Chen, and Zhao 2025), Vision-R1-7B (Huang et al. 2025), R1-VL-7B (Lu, Zhong, and Liu 2025), R1-Onevision-7B (Zhu, Li, and Wang 2025), OpenVLThinker-7B (Chen, Kumar, and Gupta 2025), ThinkLite-VL-7B (Wang et al. 2025b), and VL-Rethinker-7B (Wang et al. 2025a). All the four reasoning-focused models adopt RL as a core component to improve multimodal reasoning capabilities. Concretely, MM-Eureka-7B follows a hybrid paradigm combining supervised fine-tuning (SFT) with subsequent RL training to refine reasoning behaviors. OpenVLThinker-7B and ThinkLite-VL-7B employ iterative self-improvement pipelines, leveraging reasoning traces from earlier model outputs. Finally, MM-Eureka-7B and Vision-R1-7B further integrate custom reward mechanisms or rule-based guidance, while VL-Rethinker-7B and ThinkLite-VL-7B introduce strategies such as forced rethinking and MCTS-guided selection to promote deeper, data-efficient reasoning. Note that all the above baselines utilize Qwen2.5-VL-7B as the backbone to perform RL training, including our method Vision-G1. To contextualize the performance of our method, we also report the results from state-of-the-art large VLMs and closed-source products as a reference, *i.e.,* GPT-4o, Claude-3.5-Sonnet, Gemini-1.5-Flash, Gemini-1.5-Pro, Qwen2.5-VL-72B, InternVL2.5-78B (Chen et al. 2024c), and InternVL3-78B (Zhu et al. 2025).

Main Results

We conduct experiments on 18 benchmarks and discuss the performance on three types of visual reasoning tasks.

Evaluation on Comprehensive Visual Reasoning Tasks. In Table 1, these 7B VLMs undergone RL training can greatly surpass the backbone VLM, *i.e.,* Qwen2.5-VL-7B-Instruct. It indicates the effectiveness of RL in eliciting the visual

Models	MathVision	MathVerse	Olympiad-Bench	WeMath	DynaMath	Avg.
GPT-4o	30.6	47.8	25.9	50.6	63.7	45.7
Claude-3.5	33.5	41.2	-	-	64.8	-
Gemini-1.5 Pro	19.2	-	-	46.0	60.5	-
Qwen2.5-VL-72B	38.1	57.6	-	-	-	-
InternVL2.5-78B	32.2	51.7	11.6	39.8	19.2	28.4
InternVL3-78B	43.1	51.0	-	46.1	35.1	-
VL-Rethinker-32B	40.5	56.9	-	-	-	-
VL-Rethinker-72B	44.9	63.5	-	-	-	-
Qwen2.5-VL-7B	25.1	46.3	20.2	47.9	55.6	39.0
InternVL2.5-8B	19.7	39.5	-	23.5	-	-
MM-Eureka-7B (Qwen)	26.9	50.3	20.1	34.9	56.3	37.7
MM-Eureka-8B (Intern)	22.2	40.4	-	-	-	-
Vision-R1-7B	**32.3**	52.4	21.1	**50.5**	56.0	**42.5**
R1-VL-7B	24.7	40.0	12.1	-	45.8	-
R1-Onevision-7B	29.9	46.4	16.9	30.0	53.1	35.3
OpenVLThinker-7B	25.3	47.9	19.5	36.9	55.0	36.9
ThinkLite-VL-7B	28.1	50.7	22.3	41.6	55.9	39.7
VL-Rethinker-7B	**32.3**	**54.2**	**24.0**	41.7	57.1	41.9
Vision-G1 (ours)	31.3	51.9	23.7	45.1	**58.5**	42.1

Table 2: Benchmarking results for math reasoning tasks, where most of the baseline results are sourced from this paper (Zhu et al. 2025) and others from official papers or leaderboards.

Models	Charxiv (R/D)	ChartQA-Pro	VQA-RAD	PathVQA	SLAKE	Muir-Bench	Avg.
GPT-4o	47.1/84.4	41.7	-	-	-	68.0	-
Claude-3.5	60.2/84.3	53.7	-	-	-	-	-
Gemini-1.5 Flash	33.9/-	46.0	-	-	-	-	-
Gemini-1.5 Pro	43.3 / 72.0	-	-	-	-	-	-
Qwen2.5-VL-7B	42.7/73.5	46.7	**74.5**	65.2	76.3	39.8	57.5
MM-Eureka-7B	41.3/67.8	39.9	40.2	46.1	57.2	23.9	41.4
Vision-R1-7B	38.9/57.6	39.7	64.9	48.5	65.1	41.3	49.7
ThinkLite-VL-7B	43.8/65.0	46.2	70.5	**68.1**	**78.6**	59.0	61.0
VL-Rethinker-7B	42.8/69.3	41.5	56.2	66.1	59.7	58.3	54.1
Vision-G1 (ours)	**44.0/65.5**	**47.7**	72.1	66.7	78.3	**61.5**	**61.7**

Table 3: Results for domain-specific reasoning tasks, where most results are sourced from this paper (Zhu et al. 2025).

reasoning ability of VLMs. Among all the RL-trained methods, ThinkLite-VL performs better in four benchmarks (*i.e.,* MathVista, MMStar, LogicVista and ChartQA). ThinkLite-VL adopts a MCTS-guided sample selection method that relies on multiple iterations to measure the difficulty score. It demonstrates that proper data filtering strategy is rather useful for improving the RL training of VLMs. Benefiting from both RL training and the carefully designed data arrangement methods, our method achieves the best performance on most of the benchmarks, achieving 1.6% absolute improvement on average. Besides, our approach performs relatively well on LogicVista dataset, which contains a variety of visual logical reasoning tasks, although we have not purposely collected related datasets. It indicates that our model has better learned generalizable reasoning knowledge from other datasets, leading to such improvement. Furthermore, our approach can achieve comparable or even better performance than larger VLMs and commercial products, *e.g.,* InvernVL2.5-78B and Gemini-1.5 Flash. It further shows the promising potential of RL and well-organized data processing pipeline.

Evaluation on Math-Related Visual Reasoning Tasks. Table 2 presents the results of math-related visual reasoning tasks. First, we can see that our proposed Vision-G1 and all the RL trained baselines show significant improvement over the base model. Among all the baselines, our model Vision-G1 performs the best on average, showing the effectiveness of the math datasets we collected during the training. We also notice that Vision-R1 performs better on MahVision, MathVerse and We Math benchmarks. We argure that it is due to their focused mathematical setting, which guarantees the performance on math reasoning tasks. However, their model do not perform well in other domain datasets, *e.g.,* MMMU and MMStar. The results show that our method not only performs better than Vision-R1 on average, it can also generalize well on other datasets. It indicates our model can

Models	MathVista	MathVision	MMStar	LogicVista	ChartQA-Pro	VQA-RAD	Muir-Bench
Qwen2.5-VL-7B	65.0	25.1	62.8	42.6	46.7	**74.5**	39.8
Vision-G1 (ours)	76.1	**31.3**	**66.0**	**50.2**	47.7	72.1	**61.5**
w/o Multi-round	74.9	29.6	64.8	46.0	**48.4**	72.1	57.7
w/o Data Selection	71.5	25.3	64.2	44.0	42.9	69.7	59.4
w/o Domain Datasets	**76.3**	30.3	65.3	48.0	44.5	68.1	58.5

Table 4: Ablation study results for comprehensive visual reasoning tasks.

well balance the different visual reasoning capabilities, and serve as a better general reasoning VLM. In addition, the performance of our model on MathVista (Lu et al. 2024) is even better than OpenAI's o1 (OpenAI 2024b).

Evaluation on Domain-specific Reasoning Tasks. As shown in Table 3 containing diverse domain-specific benchmarks, the performance of RL trained VLMs even degrades a lot, compared with the backbone. Such performance degradation is more significant in the domains that the RL training data has not covered. In contrast, our method exhibits stronger robustness and better performance in these benchmarks. It further demonstrates the effectiveness of our approach as a general reasoning VLM. In our approach, we utilize influence function based data filtering method to remove the potentially harmful instances for either other training datasets or the downstream tasks. Besides, we guarantee that all the task and domain datasets should have a balanced distribution. These two strategies make our training dataset more suitable for the VLM to learn general visual reasoning abilities. Additionally, our multi-round RL training strategy with difficulty-based data selection method also ensures the stability of the knowledge learning process.

Further Analysis

In this part, we further analyze the effectiveness of our method, including the ablation study and training process analysis. More analysis experiments are in Appendix.

Ablation Study. In our method, our key designs include the multi-round RL training, data selection, and mixing multiple domain-specific datasets. In this part, we conduct the ablation study by removing one of the above design in our method. Concretely, we test the following variations of our method: (1) *w/o Multi-round* performs one-round RL training until convergence; (2) *w/o Data Selection* directly mixes up the instances from different sources without filtering and selection; (3) *w/o Domain-specific Datasets* only trains on two high-quality datasets (*i.e.,* ThinkLite and ViRL39k). As the results shown in Table 4, all the variations perform not better than our method. It indicates that all the above designs are necessary to guarantee our good performance. Besides, the variation w/o Domain-specific Datasets can achieve better performance on MathVista, but degrades a lot on most domain-specific datasets. It further proves that adding a variety of domain-specific training dataset is important for learning general reasoning ability of VLMs.

Training Process Analysis. Since our approach incorporates the multi-round RL training with proper data filtering and selection strategies, its training process will be more stable and the ability can be gradually improved. To verify the training stability and effectiveness, we collect the mean reward scores and the MathVista-Mini test scores of each 10 steps checkpoints during the two-round training process. As in the visualized results in Fig. 4, we can see that the mean reward score converges fast in the first round, and the performance on MathVista also starts to converge. Then, as we move to the next round, our difficulty-based data selection strategy will be used to construct a moderate difficult training dataset. Thus, we see a drop in the reward score in the 100th step, and then both the reward and the performance of the test set can also improve.

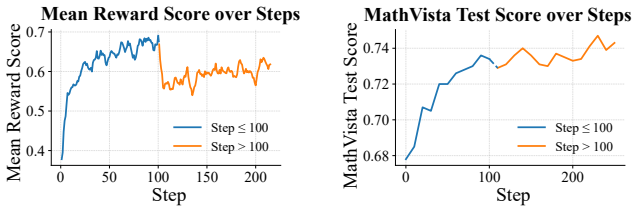

Figure 4: The mean reward scores and accuracy on MathVista_MINI during two-round RL training process.

Conclusion

In this paper, we presented Vision-G1, a general reasoning VLM trained with reinforcement learning to mimic human real-world learning via interactions with a reward model. We first constructed a large RL-ready dataset for general reasoning VLMs by assembling 46 tasks across 13 dimensions in 5 domains and unifying their data format. We then proposed an influence-function-based data filtering strategy to remove low-quality instances. After filtering, we performed multi-round RL training with GRPO, alternating difficulty-aware data selection and RL optimization to gradually improve the model's general reasoning ability. In each round, we ensured that the intermediate RL training dataset contains instances of moderate difficulty and a well-balanced category distribution. Experiments on 17 benchmarks demonstrate the effectiveness of our method, surpassing state-of-the-art baselines of similar scale and even outperforming GPT-4o and Gemini-1.5.

In future work, we plan to extend our method to more real-world tasks and scenarios, *e.g.,* video understanding and 3D perception. We will also investigate on-policy data synthesis methods that automatically generate new high-value instances to further enhance intermediate training.

References

Antol, S.; Agrawal, A.; et al. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Bai, S.; Chen, K.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chen, B.; Xu, Z.; et al. 2024a. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14455–14465.

Chen, H.; Tu, H.; et al. 2025a. SFT or RL? An Early Investigation into Training R1-Like Reasoning Large Vision-Language Models. *arXiv preprint arXiv:2504.11468*.

Chen, H.; Zhang, Y.; et al. 2023. Maybe only 0.5% data is needed: A preliminary exploration of low training data instruction tuning. *arXiv preprint arXiv:2305.09246*.

Chen, L.; Li, J.; et al. 2024b. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Chen, L.; Li, L.; et al. 2025b. R1-V: Reinforcing Super Generalization Ability in Vision-Language Models with Less Than $3. https://github.com/Deep-Agent/R1-V. Accessed: 2025-02-02.

Chen, Y.; Kumar, R.; and Gupta, A. 2025. OpenVLThinker-7B: MCTS-Guided Reasoning in Vision–Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2025*.

Chen, Z.; Wang, W.; et al. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Chen, Z.; Wu, J.; et al. 2024d. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.

Cobbe, K.; Kosaraju, V.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

DeepSeek-AI; Guo, D.; et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Du, Y.; Liu, Z.; et al. 2025. Virgo: A Preliminary Exploration on Reproducing o1-like MLLM. *arXiv preprint arXiv:2501.01904*.

Guo, D.; Yang, D.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

He, C.; Luo, R.; et al. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

He, X.; Zhang, Y.; et al. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Hendrycks, D.; Burns, C.; et al. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. *NeurIPS*.

Hu, E. J.; Shen, Y.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Hu, J.; Zhang, Y.; et al. 2025. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*.

Huang, W.; Jia, B.; et al. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Hudson, D. A.; and Manning, C. D. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. arXiv:1902.09506.

Jain, N.; Zhang, T.; et al. 2023. Llm-assisted code cleaning for training accurate code generators. *arXiv preprint arXiv:2311.14904*.

Jiang, D.; He, X.; et al. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

Jimenez, C. E.; Yang, J.; et al. 2024. SWE-bench: Can Language Models Resolve Real-world Github Issues? In *The Twelfth International Conference on Learning Representations*.

Kazemi, M.; Dikkala, N.; et al. 2024. ReMI: A Dataset for Reasoning with Multiple Images. arXiv:2406.09175.

Lau, J. J.; Gayen, S.; et al. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1): 1–10.

Li, Z.; Wu, X.; et al. 2025. A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges. arXiv:2501.02189.

Liu, B.; Zhan, L.-M.; et al. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, 1650–1654. IEEE.

Liu, H.; Li, C.; et al. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; Li, C.; et al. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, H.; Li, C.; et al. 2023b. Visual Instruction Tuning. arXiv:2304.08485.

Liu, W.; Zeng, W.; et al. 2023c. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *arXiv preprint arXiv:2312.15685*.

Liu, Z.; Sun, Z.; et al. 2025. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*.

Lu, F.; Zhong, Z.; and Liu, S. 2025. R1-VL-7B: Cross-Modal Policy Optimization for Step-wise Reasoning. In *Advances in Neural Information Processing Systems (NeurIPS) 2025*.

Lu, P.; Bansal, H.; et al. 2024. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In *International Conference on Learning Representations (ICLR)*.

Masry, A.; Do, X. L.; et al. 2022. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2263–2279.

Masry, A.; Islam, M. S.; et al. 2025. ChartQAPro: A More Diverse and Challenging Benchmark for Chart Question Answering. *arXiv preprint arXiv:2504.05506*.

Masry, A.; Kavehzadeh, P.; et al. 2023. UniChart: A Universal Vision-language Pretrained Model for Chart Comprehension and Reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 14662–14684.

Meng, F.; Du, L.; et al. 2025a. MM-Eureka: Exploring Visual Aha Moment with Rule-based Large-scale Reinforcement Learning. *arXiv preprint arXiv:2503.07365*.

Meng, F.; Du, L.; et al. 2025b. MM-Eureka: Exploring Visual Aha Moment with Rule-based Large-scale Reinforcement Learning. *arXiv preprint arXiv:2503.07365*.

Meng, X.; Chen, L.; and Zhao, W. 2025. MM-Eureka-8B: An Intern-Optimized Variant for Multimodal Reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Muennighoff, N.; Liu, Q.; et al. 2023. Octopack: Instruction tuning code large language models. *arXiv preprint arXiv:2308.07124*.

OpenAI. 2024a. GPT-4o System Card. https://openai.com/index/gpt-4o-system-card/.

OpenAI. 2024b. OpenAI o1 System Card. https://openai.com/index/openai-o1-system-card/.

Peng, Y.; Zhang, G.; et al. 2025. LMM-R1: Empowering 3B LMMs with Strong Reasoning Abilities Through Two-Stage Rule-Based RL. arXiv:2503.07536.

Pruthi, G.; Liu, F.; et al. 2020. Estimating Training Data Influence by Tracking Gradient Descent. *ArXiv*, abs/2002.08484.

Qiao, R.; Tan, Q.; et al. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*.

Qwen Team. 2024. QVQ: To See the World with Wisdom.

Shao, Z.; Wang, P.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shen, H.; Liu, P.; et al. 2025. VLM-R1: A Stable and Generalizable R1-style Large Vision-Language Model. arXiv:2504.07615.

Team, K.; Du, A.; et al. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Wang, F.; Fu, X.; et al. 2024a. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*.

Wang, H.; Qu, C.; et al. 2025a. VL-Rethinker: Incentivizing Self-Reflection of Vision-Language Models with Reinforcement Learning. *arXiv preprint arXiv:2504.08837*.

Wang, K.; Pan, J.; et al. 2024b. Measuring Multimodal Mathematical Reasoning with MATH-Vision Dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Wang, X.; Yang, Z.; et al. 2025b. SoTA with Less: MCTS-Guided Sample Selection for Data-Efficient Visual Reasoning Self-Improvement. *arXiv preprint arXiv:2504.07934*.

Wang, Z.; Xia, M.; et al. 2024c. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37: 113569–113697.

Wei, J.; Wang, X.; et al. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

Xia, M.; Malladi, S.; et al. 2024. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*.

Xiao, Y.; Sun, E.; et al. 2024. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*.

Xu, G.; Jin, P.; et al. 2024. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

Xu, G.; Jin, P.; et al. 2025. LLaVA-CoT: Let Vision Language Models Reason Step-by-Step. arXiv:2411.10440.

Xu, Z.; Du, S.; et al. 2023. Chartbench: A benchmark for complex visual reasoning in charts. *arXiv preprint arXiv:2312.15915*.

Yang, Y.; He, X.; et al. 2025. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*.

Yao, Y.; Yu, T.; et al. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800*.

Yu, Q.; Zhang, Z.; et al. 2025. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. arXiv:2503.14476.

Yuan, Y.; Yu, Q.; et al. 2025. VAPO: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*.

Yue, X.; Ni, Y.; et al. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9556–9567.

Yue, X.; Zheng, T.; et al. 2024b. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*.

Zhan, Y.; Zhu, Y.; et al. 2025. Vision-R1: Evolving Human-Free Alignment in Large Vision-Language Models via Vision-Guided Reinforcement Learning. *arXiv preprint arXiv:2503.18013*.

Zhang, R.; Jiang, D.; et al. 2024a. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, 169–186. Springer.

Zhang, R.; Zhang, B.; et al. 2024b. Improve Vision Language Model Chain-of-thought Reasoning. arXiv:2410.16198.

Zhao, B.; Zong, Y.; et al. 2024. Benchmarking Multi-Image Understanding in Vision and Language Models: Perception, Knowledge, Reasoning, and Multi-Hop Reasoning. arXiv:2406.12742.

Zhou, C.; Liu, P.; et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Zhu, J.; Wang, W.; et al. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*.

Zhu, X.; Li, H.; and Wang, J. 2025. R1-Onevision-7B: Formalizing Cross-Modal Reasoning for Generalization. In *Annual Meeting of the Association for Computational Linguistics (ACL) 2025*.

Zhuo, T. Y.; Zebaze, A.; et al. 2024. Astraios: Parameter-Efficient Instruction Tuning Code Large Language Models. *arXiv preprint arXiv:2401.00788*.

Zou, C.; Guo, X.; et al. 2024. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836*.