Multimodal Mathematical Reasoning with Diverse Solving Perspective

Anonymous ACL submission

Abstract

Recent progress in large-scale reinforcement learning (RL) has notably enhanced the reasoning capabilities of large language models (LLMs), especially in mathematical domains. However, current multimodal LLMs (MLLMs) for mathematical reasoning often rely on one-to-one image-text pairs and single-solution supervision, overlooking the diversity of valid reasoning perspectives and internal reflections. In this work, we introduce MathV-DP, a novel dataset that captures multiple diverse solution trajectories for each image-question pair, fostering richer reasoning supervision. We further propose Qwen-VL-DP, a model built upon Qwen-VL, fine-tuned with supervised learning and enhanced via group relative policy optimization (GRPO), a rule-based RL approach that integrates correctness discrimination and diversity-aware reward functions. Our method emphasizes learning from varied reasoning perspectives and distinguishing between correct yet distinct solutions. Extensive experiments on the MathVista's minitest and Math-V benchmarks demonstrate that Qwen-VL-DP significantly outperforms prior base MLLMs in both accuracy and generative diversity, highlighting the importance of incorporating diverse perspectives and reflective reasoning in multimodal mathematical reasoning. We will make our data and model public available.

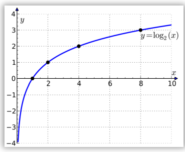
1 Introduction

Large language models (LLMs) have demonstrated remarkable abilities in reasoning tasks (Wei et al., 2022; Wang et al., 2023; Zhou et al., 2023). This has spurred significant interest in their application to solving math problems described in natural language (Luo et al., 2023; Yue et al., 2023b; Gou et al., 2023; Jiang et al., 2023). Meanwhile, a more challenging direction involves multimodal mathematical reasoning (Lu et al., 2023), where models must interpret various types of images and apply advanced logical skills to address mathematical questions with visual components. Open-source multimodal large language models (MLLMs), such as LLaVA (Liu et al., 2023) and Qwen-VL (Bai et al., 2023), have achieved strong results on visual question answering benchmarks (Guo et al., 2023). However, when it comes to intricate mathematical problems that require visual understanding, these models still lag behind close-source counterparts like GPT-4V and Gemini (OpenAI-b; Google).

Humans frequently engage in intuitive chain-of-thought (CoT) processes to address complex reasoning tasks (Ericsson and Simon, 1980). Recent research (Wei et al., 2022) has demonstrated that LLMs are capable of exhibiting similar CoT reasoning. By employing straightforward prompting strategies or fine-tuning methods (Wang et al., 2023; Hsieh et al., 2023), CoT can both boost the reasoning abilities of LLMs and increase transparency in their decision-making procedures. Notably, recent progresses, such as OpenAI o1 (OpenAI-c), have enabled LLMs to generate more elaborate internal CoT sequences. Despite these successes in natural language contexts, adapting CoT approaches for multimodal tasks remains fully unexplored. In contrast to the rich supply of text-centric CoT data used during language model training, there is a marked shortage of multimodal CoT datasets within predominantly text-based online resources (Dai et al., 2024). This scarcity constrains the development and reasoning capacity of MLLMs.

Recent advancements in large-scale reinforcement learning (RL) (Sutton et al., 1998) have significantly enhanced the reasoning capacity of LLMs especially within mathematical reasoning tasks. o1 (OpenAI-c) and DeepSeek-R1 (Guo et al., 2025) illustrate that extensive RL applied during post-training can lead to substantial gains in complex reasoning performance, in some instances surpassing outcomes achieved via supervised fine-tuning

Question: The derivative of y at x=6 is __ that at x=8.



✗ **Solution 1:** (Confusing Function Value with Derivative) Consider the value of the function. log$_2$(6)≈2.584. log$_2$(8)=3.We can see that 3>2.584, so conclude the answer is smaller than.

✓ **Solution 2:** (Using the Derivative Formula) Recall the formula. For y=log$_2$(x), the derivative is: dy/dx=1/xln2. Plug in the values. At x=6: 1/6ln2. At x=8: 1/8ln2. Both denominators are positive. Since 6<8, we have 1/6ln2 > 1/8ln2. It is clear that as x increases, the denominator increases and the derivative (the value) gets smaller. The derivative at x=6 is larger than that at x=8.

✓ **Solution 3:** (Graphical/Intuitive Reasoning) Look at the graph, the curve y=log$_2$(x) is plotted and shows a shape that rises quickly for small x, but becomes flatter as x increases. The derivative at any point corresponds to the slope of the curve at that point. Near x=6, the curve is a bit steeper. x=8, the curve is flatter (the tangent is closer to horizontal). Since the curve gets flatter as x increases, the slope at x=6 must be greater than at x=8.

Figure 1: An multimodal mathematical reasoning example with alternative solutions that reaches the final answer. Existing open-source image instruction datasets containing limited solution per image-question, do not fully exploit diverse solution with reflection to enhance the multimodal mathematical reasoning capabilities of MLLMs.

(SFT) (Radford et al., 2019). There has been growing interest within the research community to adapt the rule-based RL used in DeepSeek-R1 to multimodal scenarios (Chen et al., 2025; Yang et al., 2025). These works just explore using final answer and thinking format of image instruction dataset as reward signal.

Furthermore, most existing MLLMs focus on pre-training and post-training by using one-to-one image-text data to improve the final answer accuracy on mathematical reasoning but neglect diverse perspective of internal thought. As shown in Figure 1, for an image-question pair, there are usually multiple reasonable inference solutions to reach the final correct answer. Constrained by limited thinking perspectives tend to derive wrong solution and answer. Existing open-source image instruction datasets for fine-tuning or reinforcement learning, containing limited solution per image-question, do not fully exploit diverse solution with reflection to enhance the multimodal mathematical reasoning capabilities of MLLMs.

To bridge the gap, we construct MathV-DP dataset involving a variety of solutions for image-question corresponding to a single thought solution, and train the model Qwen-VL-DP based on the Qwen-VL-7B (Bai et al., 2025; Wang et al., 2024c) through supervised fine-tuning and group relative policy optimization (GRPO) (Shao et al., 2024) as rule-based reinforcement learning. In addition, the discrimination of diverse correct solutions and the preference for different correct and incorrect solutions are introduced in the reward function. Experiments on MathVista's minitest (Lu et al., 2023) and Math-V (Wang et al., 2024a) show that learning the correctness, diverse skills and reasoning trajectories from multiple solution perspectives significantly improves the accuracy and generation diversity of base MLLMs on multimodal mathematical reasoning.

2 Related Works

2.1 Multimodal Reasoning

The progress of MLLMs has significantly advanced research in multimodal reasoning (Chen et al., 2024; You et al., 2023). A widely adopted strategy involves augmenting existing question-answer datasets in specialized domains to further fine-tune MLLMs. For answer enhancement, rationales have been either human-authored (Zhang et al., 2023) or extracted from leading LLMs (Wang et al., 2024b; Lin et al., 2023a; Chen and Feng, 2023; Li et al., 2024). Furthermore, VPD (Hu et al., 2023) introduced a method for converting programmatic answer representations into natural language explanations. On the question side, DDCoT (Zheng et al., 2023) employed LLMs to decompose complex queries into simpler sub-questions. Math-LLaVA (Shi et al., 2024) explored raw visual information presented in images to construct more questions. To provide a more comprehensive assessment of MLLM multimodal reasoning, several benchmarks have emerged: MathVista (Lu et al., 2023), and Math-V (Wang et al., 2024a) address diverse mathematical reasoning tasks, while MMMU (Yue et al., 2023a) spans multiple disciplines. Despite these progresses, open-source MLLMs still exhibit substantial room for improvement in complex multimodal reasoning scenarios.

2.2 Reinforcement Learning

With the advent of LLMs (Brown et al., 2020; Radford et al., 2018), reinforcement learning from human feedback (RLHF) (Bai et al., 2022) has emerged as an essential strategy for model fine-tuning, utilizing human-annotated preference data. RLHF commonly incorporates optimization methods like proximal policy optimization (PPO) (Schulman et al., 2017) and direct preference optimization (DPO) (Rafailov et al., 2023), facilitating improved response alignment, coherence, and utility in generated outputs.

Recently, there has been a growing interest in leveraging RL to enhance the reasoning abilities of LLMs (Team et al., 2025; Guo et al., 2025; Shao et al., 2024; Luong et al., 2024), particularly within the scope of mathematical reasoning. The central approach involves designing reward functions or evaluative models that preferentially reinforce high-quality reasoning steps and discourage inadequate reasoning, thereby steering the optimization process toward more organized and comprehensible reasoning patterns through RL techniques. For instance, ReST-MCTS (Zhang et al., 2024) utilizes a process reward model (PRM) to assess the correctness of individual reasoning steps within solution paths. Moreover, recent research indicates that even straightforward rule-based, outcome-level reward functions can serve as robust and informative signals during RL, as demonstrated by DeepSeek-R1 (Guo et al., 2025). DeepSeek-R1 incorporates group relative policy optimization (GRPO) (Shao et al., 2024) combined with outcome-based reward assessments, effectively advancing the reasoning proficiency of LLMs. In this work, we focus on further enhancing the reasoning capabilities of MLLMs through reinforcement learning.

3 Method

Our proposed method is composed of two components: (1) bootstrapping a substantial set of both positive and negative chain-of-thought (CoT) solutions with reflection for collected multimodal mathematical question-CoT; and (2) leveraging these new sampled positive solutions, pairs of different positive solutions and pairs of positive-negative solutions to perform post-training on the underlying diverse rationales and to facilitate learning discrimination and preference from identified pairs. Through the data synthesis and post-training, the MLLM is progressively improved from an initial single solving perspective to a diverse state. The overall framework is depicted in Figure 2.

3.1 Data Synthesis

In vision-language reasoning tasks, given an image I and a corresponding question q, an MLLM is expected to perform joint reasoning over both modalities to generate a rationale r, followed by deriving a final answer a. However, constructing large-scale datasets comprising high-quality (I, q, r, a) remains a significant challenge, primarily due to the scarcity of well-annotated rationale data. This data bottleneck hinders the post-training enhancement of MLLM reasoning capabilities. Although MLLMs possess a rudimentary ability for CoT reasoning and self-reflection, leveraging them to generate diverse and high-quality (I, q, r, a) samples from existing multimodal mathematical datasets is difficult. Recent advancements in language models, such as DeepSeek-R1, demonstrate strong capabilities in producing coherent, reflective reasoning across extended textual contexts. Formal languages, characterized by strict syntactic and semantic rules, provide a structured representation that eliminates ambiguity and enforces logical consistency. When visual content is described using formal language, it enables language models to see and reason over image elements more effectively. In our work, we utilize DeepSeek-R1 (Guo et al., 2025) to synthesize diverse detailed reasoning chains on samples from the MultiMath-300K dataset (Peng et al., 2024). This facilitates the construction of a richer and more diverse set of cross-modal mathematical reasoning samples, culminating in our proposed 40K MathV-DP dataset. The data generation pipeline is illustrated on the left side of Figure 2.

Data Source. We adopt MultiMath-300K (Peng et al., 2024) as the primary data source for our data synthesis. This dataset is a large-scale, multimodal, multilingual, multi-level, and multi-step mathematical reasoning benchmark, encompassing a wide range of K-12 level problems. It spans nearly the entire K-12 curriculum, covering a broad spectrum of mathematical domains, including arithmetic, algebra, geometry, functions, algorithms, and more. Compared to existing multimodal mathematics datasets (e.g., Geo170K (Gao et al., 2023) and MathV360K (Shi et al., 2024)), the problems in MultiMath-300K are newly curated from the real world and do not overlap with those in previously released datasets lacking high-quality CoT annotations or containing only final answers. Each instance is paired with a descriptive image caption to support vision-language alignment, as well as a detailed step-by-step solution. The availability of formal visual descriptions and CoT annotations in MultiMath-300K with single solution per sample makes it particularly well-suited as seed data for synthesizing diverse solutions from multiple perspectives. Specifically, we randomly selected 10K samples from them as seed data $\mathcal{D}$.

Diverse Solutions Construction. Given an image,

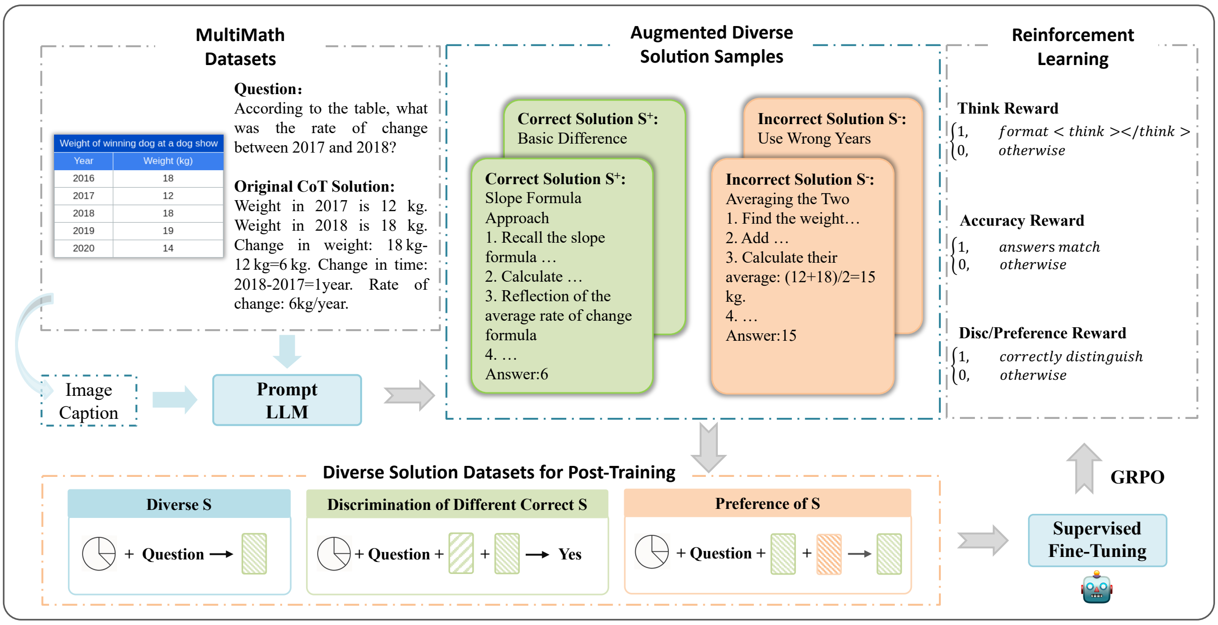

MultiMath Datasets
Question:
According to the table, what was the rate of change between 2017 and 2018?
Weight of winning dog at a dog show
Year
Weight (kg)
2016
18
2017
12
2018
18
2019
19
2020
14
Original CoT Solution:
Weight in 2017 is 12 kg. Weight in 2018 is 18 kg. Change in weight: 18 kg-12 kg=6 kg. Change in time: 2018-2017=1year. Rate of change: 6kg/year.
Image Caption
Prompt LLM
Augmented Diverse Solution Samples
Correct Solution S+: Basic Difference
Correct Solution S+: Slope Formula Approach
1. Recall the slope formula …
2. Calculate …
3. Reflection of the average rate of change formula
4. …
Answer:6
Incorrect Solution S-: Use Wrong Years
Incorrect Solution S-: Averaging the Two
1. Find the weight…
2. Add …
3. Calculate their average: (12+18)/2=15 kg.
4. …
Answer:15
Reinforcement Learning
Think Reward
1, format < think ></think >
0, otherwise
Accuracy Reward
1, answers match
0, otherwise
Disc/Preference Reward
1, correctly distinguish
0, otherwise
GRPO
Diverse Solution Datasets for Post-Training
Diverse S
+ Question
Discrimination of Different Correct S
+ Question +
+
Yes
Preference of S
+ Question +
+
Supervised Fine-Tuning

Figure 2: The overall flowchart of the proposed multimodal question-solution data synthesis and post-training. Post-training consists of supervised fine-tuning and rule-based reinforcement learning (GRPO) to learn diverse and reflection reasoning manner.

we prompt large language reasoning model (i.e., DeepSeek-R1) with its formal dense caption, question and limited original solution to construct more diverse CoT data with reflection. The prompt for generating new solutions s is shown in Figure 3, guiding the model to identify the whole objective and provide a general idea of the plan, propose corrections or alternative reasoning paths, verify consistency between its intermediate reasoning and the final answer. Two correct solutions and two incorrect solutions that differ from each other are generated at once for each source sample to reduce expenses. They are organized into three formats to constitute MathV-DP dataset involving CoT with reflection thinking, discrimination of different correct solutions and preference of solutions.

The correct solution with reflection is first taken out separately with the original image and question. The rationale before the final answer in each solution is wrapped with <*think*> and </*think*> tags as r_{think} to form a new set $\mathcal{D}_s^+$ totaling 20K:

$$\mathcal{D}_s^+ = \{(I_i, q_i, r_{think}, a_i)\}_{i=1}^{|\mathcal{D}|} \tag{1}$$

The generated different correct solutions are then concatenated with the instruction Ins_1 (i.e., *"Are the solution perspectives of the two solutions dissimilar?"*) to form set $\mathcal{D}_d$ totaling 10K used for the calculation of discrimination reward during RL:

$$\mathcal{D}_d = \left\{\left(I_i, q_i, s_{i1}^+, s_{i2}^+, Ins_1, 1\right)\right\}_{i=1}^{|\mathcal{D}|} \tag{2}$$

For the data format of correctness preference and future perference reward calculation, one of each of the correct and incorrect solutions is randomly selected and both are concatenated together as a pair in a random back-and-forth order to construct set $\mathcal{D}_p$ totaling 10K. Instruction Ins_2 is *"Is the former/later solution the correct one?"*:

$$\mathcal{D}_p = \left\{\left(I_i, q_i, s_i^+, s_i^-, Ins_2, 1\right)\right\}_{i=1}^{|\mathcal{D}|} \tag{3}$$

Prompt-Solutions Generation:
[Role] You are a math expert.
[Image Description] formal image description
[Original Question] question
[Given Solution] original solution
[Task] Please change the given solution to two solutions that are different and incorrect. Then change the given solution to two solutions that have different solution paths but the same final right answer.
[Requirement] For each solution, please involve complete and detailed seeking thought process with planning, reflection, and verification. Please split each whole solution with '/******/.' The solutions should be coherent and independent, and contain no information about the correct given solution.'''

Figure 3: The prompt template used in our DeepSeek-R1 API for generating additional solutions with reflection for each input image description, question and original CoT solution.

3.2 Post-Training

To improve the multimodal mathematical reasoning capabilities of MLLMs, we propose a two-stage post-training framework comprising supervised fine-tuning followed by rule-based reinforcement learning. In this pipeline, supervised fine-tuning serves to stabilize the model's reasoning ability and learn diverse solving process with reflection, while the subsequent reinforcement learning phase promotes better generalization, preference of solution correctness and diversity in multimodal mathematical reasoning task.

3.2.1 Supervised Fine-Tuning

Specifically, we utilize $\mathcal{D}_s^+$ with diverse solution perspectives during the supervised fine-tuning stage to guide the model $\mathcal{M}$ toward generating coherent and diverse reasoning chains with a negative log-likelihood objective:

$$\mathcal{L}_{\text{SFT}} = -\sum_{(I,q,r,a)\sim\mathcal{D}_s^+} \log \mathcal{M}(r, a \mid q, I) \quad (4)$$

SFT not only aligns the model's outputs with desired formats but also encourages the emergence of more sophisticated multimodal mathematical reasoning reflection behaviors. This establishes a robust foundation for the subsequent RL phase, where rule-based feedback is employed to further refine the model's reasoning abilities.

3.2.2 Rule-Based Reinforcement Learning

Building upon the model fine-tuned via SFT we further optimize its structured reasoning capabilities, output validity and diversity of solutions through a rule-based reinforcement learning framework. In particular, we design three reward functions and employ group relative policy optimization (GRPO) (Shao et al., 2024) for policy updates.

Accuracy Reward. The accuracy reward function assesses the correctness of the MLLM's final output by extracting the predicted answer using regular expressions and comparing it against the ground truth. We regard multimodal mathematical reasoning as deterministic tasks, the model is required to present the final answer in a predefined format to facilitate consistent and rule-based evaluation.

Think Format Reward. To enforce the explicit presence of a reasoning process, the format-based reward function mandates that the MLLM's rationale be encapsulated within predefined delimiters, i.e., <*think*> and </*think*>. Regularization is used to verify the existence and correct ordering of these markers, thereby ensuring adherence to the required output structure.

Discrimination and Preference Reward. The discrimination/preference reward function can be viewed as a binary classification task. It is used to evaluate whether the MLLM correctly distinguishes the diversity of different solutions and whether it prefers the correct solution. This reward signal facilitates the model to learn the different perspectives of the solutions and the correctness preference.

Group Relative Policy Optimization. To ensure stable training with both consistent policy updates and informative reward signals, we adopt group relative policy optimization (GRPO) as our reinforcement learning algorithm. For each token in the generated sequence, GRPO computes the log-likelihoods under the current policy $\pi(\theta)$ and a reference policy. The ratio between these probabilities is then calculated and clipped within the interval $[1-\epsilon, 1+\epsilon]$ to mitigate the risk of overly aggressive updates. The reward, normalized to serve as an advantage estimate, is subsequently incorporated into a proximal policy optimization (PPO) objective function:

$$\mathcal{L}_{\text{clip}} = -\mathbb{E}[\min(\text{ratio}_t \cdot Ad_t, \text{clipratio}_t \cdot Ad_t)], \quad (5)$$

where Ad_t represents the advantage estimate, quantifying the relative improvement of the chosen action over the expected value under the reference policy. To further constrain the updated policy from deviating excessively from the reference distribution, a Kullback–Leibler (KL) divergence term is incorporated into the objective, scaled by a coefficient β. The total loss function is defined as:

$$\begin{aligned}\mathcal{L}_{\text{RL}}(\theta) = -\mathbb{E}[\,&\min\left(\text{ratio}_t \cdot Ad_t, \text{clipratio}_t \cdot Ad_t\right) \\ &- \beta \cdot \text{KL}\left(\pi_\theta(y|x), \pi_{\text{ref}}(y|x)\right)]\end{aligned} \quad (6)$$

GRPO employs a clipping strategy that effectively mitigates drastic changes in the policy, while the incorporation of KL regularization enforces proximity between the updated and reference policies. This dual mechanism enables stable and efficient integration of rule-based rewards, preserving training robustness throughout the optimization process.

4 Experiments

4.1 Model and Implementation

We utilize the Qwen2-VL (Wang et al., 2024c) and Qwen2.5-VL (Bai et al., 2025) series as our base-

Model	**MathVista**												
	ALL	FQA	GPS	MWP	TQA	VQA	ALG	ARI	GEO	LOG	NUM	SCI	STA
Heuristics Baselines													
Random Chance	17.9	18.2	21.6	3.8	19.6	26.3	21.7	14.7	20.1	13.5	8.3	17.2	16.3
Frequent Guess (Lu et al., 2023)	26.3	22.7	34.1	20.4	31.0	24.6	33.1	18.7	31.4	24.3	19.4	32.0	20.9
Human	60.3	59.7	48.4	73.0	63.2	55.9	50.9	59.2	51.4	40.7	53.8	64.9	63.9
Close-Source Multimodal Large Language Models (MLLMs)													
Gemini 1.0 Nano 2 (Team et al., 2023)	30.6	28.6	23.6	30.6	41.8	31.8	27.1	29.8	26.8	10.8	20.8	40.2	33.5
Qwen-VL-Plus (Bai et al., 2023)	43.3	**54.6**	38.5	31.2	55.1	34.1	39.1	32.0	39.3	18.9	26.4	59.0	56.1
Gemini 1.0 Pro (Team et al., 2023)	45.2	47.6	40.4	39.2	61.4	**39.1**	45.2	38.8	41.0	10.8	**32.6**	54.9	**56.8**
Claude 3 Haiku (Anthropic, 2024)	46.4	-	-	-	-	-	-	-	-	-	-	-	-
GPT-4V (OpenAI-b)	49.9	43.1	**50.5**	**57.5**	**65.2**	38.0	**53.0**	**49.0**	**51.0**	**21.6**	20.1	**63.1**	55.8
GPT-4o (OpenAI-a)	63.8	-	-	-	-	-	-	-	-	-	-	-	-
OpenAI o1 (OpenAI-c)	**73.9**	-	-	-	-	-	-	-	-	-	-	-	-
Open-Source Multimodal Large Language Models (MLLMs)													
miniGPT4-7B (Zhu et al., 2023)	23.1	18.6	26.0	13.4	30.4	30.2	28.1	21.0	24.7	16.2	16.7	25.4	17.9
InstructBLIP-7B (Dai et al., 2024)	25.3	23.1	20.7	18.3	32.3	35.2	21.8	27.1	20.7	18.9	20.4	33.0	23.1
LLaVA-13B (Liu et al., 2023)	26.1	26.8	29.3	16.1	32.3	26.3	27.3	20.1	28.8	24.3	18.3	37.3	25.1
SPHINX-V1-13B (Lin et al., 2023b)	27.5	23.4	23.1	21.5	39.9	34.1	25.6	28.1	23.4	16.2	17.4	40.2	23.6
LLaVA-1.5-13B (Liu et al., 2024)	27.7	23.8	22.7	18.3	40.5	30.2	25.3	26.4	22.8	21.6	26.4	35.3	23.6
OmniLMM-12B (OpenBMB, 2024)	34.9	45.0	17.8	26.9	44.9	39.1	23.1	32.3	20.9	18.9	27.8	45.9	44.2
SPHINX-V2-13B (Lin et al., 2023b)	36.7	54.6	16.4	23.1	41.8	43.0	20.6	33.4	17.6	24.3	21.5	43.4	51.5
G-LLaVA-13B (Gao et al., 2023)	-	-	56.7	-	-	-	-	-	-	-	-	-	-
Math-LLaVA (Shi et al., 2024)	46.6	37.2	57.7	56.5	51.3	33.5	53	40.2	56.5	16.2	33.3	49.2	43.9
Math-PUMA-7B (Zhuang et al., 2025)	47.9	-	48.1	-	-	-	-	-	47.3	-	-	-	-
Multimath-7B (Peng et al., 2024)	50.0	-	66.8	61.8	-	-	-	-	-	-	-	-	-
Mulberry-7B (Yao et al., 2024)	63.1	-	-	-	-	-	-	-	-	-	-	-	-
Qwen2-VL-7B (Wang et al., 2024c)	57.6	65.1	41.8	66.1	60.1	53.7	44.5	56.4	43.1	24.3	39.6	63.1	69.4
Qwen2.5-VL-7B (Bai et al., 2025)	68.2	72.5	66.8	76.9	66.7	54.3	70.1	68.7	66.9	26.9	43.0	65.7	76.1
LLaVA-1.5-DP	42.2	32.3	56.2	51.6	45.6	39.7	43.8	41.6	46.0	15.4	38.9	46.7	40.9
Qwen2-VL-DP	60.9	70.7	56.4	69.8	64.6	48.7	50.9	60.4	47.2	25.4	40.3	65.6	71.1
Qwen2.5-VL-DP	**70.4**	**72.8**	**72.6**	**77.2**	**68.5**	**54.5**	**71.1**	**69.6**	**69.3**	**27.0**	**43.1**	**66.9**	**77.2**

Table 1: Comparison with baselines on the testmini set of MathVista benchmark. Baseline results are obtained from Lu et al. (2023). The best results in both the close-source and open-source MLLMs are in bold. MathVista is divided in two ways: task type or mathematical skill, and we report the accuracy under each subset.

line architectures and focus our evaluation on the 7B parameter scale to assess the effectiveness of our proposed method. Both the projection layer and the language model parameters are trainable. Supervised fine-tuning stage is performed with a batch size of 16, a learning rate of 2e-5 over 1 epoch. During the reinforcement learning stage, we generate 4 rollouts per query with a sampling temperature of 1.0. The maximum sequence length is set to 1024 to ensure the model has sufficient capacity to produce complete reasoning solution. Both the policy and reference models are initialized from the same base model, with the reference model held frozen during RL training. The policy model is fine-tuned using a learning rate of 1e-6 and a batch size of 4. The KL divergence regularization coefficient β in Eq. 6 is set to 0.04 by default. All experiments are conducted on NVIDIA H100 GPU with 80GB of memory.

4.2 Evaluation and Metrics

We assess our model's performance in a zero-shot setting similar to other models using the minitest subset of the MathVista benchmark (Lu et al., 2023). This subset comprises 1,000 items, including 540 multiple-choice problems and 460 free-response questions requiring answers in the form of integers, floating-point numbers, or lists. MathVista is designed to comprehensively evaluate the multimodal mathematical capabilities of MLLMs, covering diverse reasoning categories such as algebraic (ALG), arithmetic (ARI), geometric (GEO), logical (LOG), numeric commonsense (NUM), scientific (SCI), and statistical reasoning

Model	**Math-V**																
	ALL	Alg	AnaG	Ari	CG	Comb	Cnt	DG	GT	Log	Angle	Area	Len	SG	Sta	Topo	TG
Heuristics Baselines																	
Human	68.8	55.1	78.6	99.6	98.4	43.5	98.5	91.3	62.2	61.3	33.5	47.2	73.5	87.3	93.1	99.8	69.0
Close-Source Multimodal Large Langugae Models (MLLMs)																	
Qwen-VL-Plus (Bai et al., 2023)	10.7	11.3	17.9	14.3	12.7	4.8	10.5	15.4	8.9	14.3	11.6	6.4	10.0	14.3	6.9	8.7	11.3
Qwen-VL-Max (Bai et al., 2023)	15.6	10.7	19.1	20.0	16.9	12.5	17.9	16.4	12.2	21.0	13.3	14.2	19.8	11.5	20.7	13.0	17.3
Gemini Pro (Team et al., 2023)	17.7	15.1	10.7	20.7	20.1	11.9	7.5	20.2	21.1	16.8	19.1	19.0	20.0	14.3	13.8	17.4	20.8
GPT-4V (OpenAI-b)	22.8	27.3	32.1	35.7	21.1	16.7	13.4	22.1	14.4	16.8	**22.0**	22.2	20.9	23.8	24.1	21.7	**25.6**
GPT-4o (OpenAI-a)	**30.4**	**42.0**	**39.3**	**49.3**	**28.9**	**25.6**	**22.4**	**24.0**	**23.3**	**29.4**	17.3	**29.8**	**30.1**	**29.1**	**44.8**	**34.8**	17.9
Open-Source Multimodal Large Langugae Models (MLLMs)																	
LLaVA-1.5-13B (Liu et al., 2024)	11.1	7.0	14.3	14.3	9.1	6.6	6.0	13.5	5.6	13.5	10.4	12.6	14.7	11.5	13.8	13.0	10.7
Math-LLaVA (Shi et al., 2024)	15.7	9.0	20.2	15.7	18.2	10.1	10.5	16.4	14.4	16.0	20.2	18.4	17.6	9.4	24.1	**21.7**	17.9
Qwen2-VL-7B (Wang et al., 2024c)	16.3	11.3	24.9	15.7	16.9	10.1	11.9	16.4	15.6	19.3	22.5	16.4	22.5	14.3	17.2	4.4	20.8
Qwen2.5-VL-7B (Bai et al., 2025)	25.0	22.0	29.8	32.1	19.5	18.5	16.4	22.1	11.1	25.2	**29.3**	27.6	28.5	22.9	34.5	17.4	22.0
Qwen2-VL-DP	17.7	15.2	20.8	20.8	20.2	12.0	7.9	20.3	21.2	16.9	19.2	19.1	23.2	14.4	13.9	17.5	20.9
Qwen2.5-VL-DP	**26.9**	**23.3**	**30.8**	**32.2**	**20.6**	**27.3**	**17.4**	**23.9**	**22.9**	**28.6**	28.9	**30.9**	**28.8**	**28.7**	**37.9**	18.4	**23.2**

Table 2: Performance Comparison on the Math-V benchmark with the accuracy metric across various mathmatical subjects. Baseline results are obtained from Wang et al. (2024a). The best results in both the close-source and open-source MLLMs are in bold.

(STA). Additionally, its questions are distributed across various subtypes, including Figure Question Answering (FQA), Geometry Problem Solving (GPS), Math Word Problem (MWP), Textbook Question Answering (TQA), and Visual Question Answering (VQA). For evaluation, we leverage GPT-4 (OpenAI-a) to extract final answers or selected choices from model responses in a few-shot manner (Lu et al., 2023) and compute accuracy by verifying the correspondence between predicted and grounded answers. In addition, we perform evaluations on Math-V (Wang et al., 2024a). Math-V contains 3,040 visual-context math problems curated from authentic math competitions.

Accuracy evaluation mainly depends on the final answer of the MLLM output, we also use the *effective semantic diversity* metric (Shypula et al., 2025) to assess the diversity of the MLLM's output solutions. For each input, model generates K responses $G_i = \{g_i^1, g_i^2, \ldots, g_i^K\}$. We then adopt the following pairwise diversity score:

$$\text{Div}_{\text{pair}}\left(G_i\right) = \frac{1}{\binom{K}{2}} \sum_{1 \leq j < k \leq K} d_{\text{sem}}\left(g_i^j, g_i^k\right), \tag{7}$$

where d_{sem} is semantic distance function. It is obtained by Sentence Transformer (Reimers and Gurevych, 2019), which is 1 if semantically dissimilar and 0 otherwise. This pairwise evaluation strategy incorporates normalization over the total number of candidate pairs, thereby ensuring robustness against fluctuations in the number of valid outputs generated for different prompts. The overall diversity of a model on the benchmark is then computed by averaging all pairwise diversity scores.

5 Results and Analysis

5.1 Main Comparison on Accuracy

We compare Qwen-VL-DP with other MLLMs on the minitest split of the MathVista benchmark in Table 1. As shown in the table, open-source MLLMs such as instructBLIP (Dai et al., 2024) and LLaVA-1.5 (Liu et al., 2023) have poor performance in multimodal mathematics, with overall accuracy lower than 30%. Compared to the base model, Qwen2.5-VL-7B, with superior multimodal mathematical ability, Qwen2.5-VL-DP achieves 70.4% overall accuracy with a improvement of 2.2%. LLaVA-1.5-DP also obtains improvement of 14.5% compared with base model LLaVA-1.5-13B. We also conducted 10 independent inference runs on Qwen2.5-VL-7B and Qwen2.5-VL-DP, observing an average improvement of 2.5% (±0.9%). The 95% confidence interval for the performance gain is (1.94%, 3.06%). More surprisingly, the proposed Qwen2.5-VL-DP model outperforms close-source models GPT-4V and GPT-4o (OpenAI-b), even achieving comparable performance to OpenAI o1 (OpenAI-c), the most powerful close-source MLLMs with the ability of detailed thinking. The results on Math-V are shown in Table 2. Qwen2.5-VL-DP demonstrates substantial performance gains over its base model, narrowing the gap with state-of-the-art models such as GPT-4V and GPT-4o. The excellent performance of Qwen-VL-DP indicates that the high-quality data synthesis of solutions with diverse perspective and reflection is effective in improving MLLM's multimodal mathematical reasoning capabilities and performance.

5.2 Comparision on Generation Diversity

The proposed Qwen-VL-DP model has demonstrated exceptional performance in multimodal mathematical reasoning task. To assess its capability of generation diversity, we conduct evaluation experiments using effective semantic diversity metric on MathVista's minitest subset. For each input sample, the number of generated responses K is taken as 3, 5, and 10 to calculate the corresponding pairwise diversity score for final averaging. Table 3 presents comparison of the effective semantic diversity among the Qwen-VL base model, the supervised fine-tuned model, the model tuned using only GRPO, and the post-training model after two stages using MathV-DP data. The results indicate that either supervised fine-tuning or reinforcement learning on MLLM using solution data with diverse perspectives can enhance the generative diversity of the base model. Through our synthesis of MathV-DP and proposed post-training, MLLM can further enhance the accuracy performance of multimodal mathematical reasoning while improving the diversity of output responses. The reason is that Qwen-VL-DP has learnt diverse solution perspectives after supervised fine-tuning and further learnt the discriminative and preference of different solutions after rule-based reinforcement learning.

Model	Diver@3	Diver@5	Diver@10
Qwen2-VL-7B	27.64	30.18	31.33
Qwen2-VL-SFT	33.72	35.63	35.75
Qwen2-VL-GRPO	35.05	36.08	37.11
Qwen2-VL-DP	37.48	38.97	39.16
Qwen2.5-VL-7B	33.29	34.76	36.89
Qwen2.5-VL-SFT	39.46	39.78	39.81
Qwen2.5-VL-GRPO	39.02	39.49	39.73
Qwen2.5-VL-DP	**40.42**	**41.44**	**41.58**

Table 3: Effective semantic diversity scores for Qwen-VL models evaluated in our experiments.

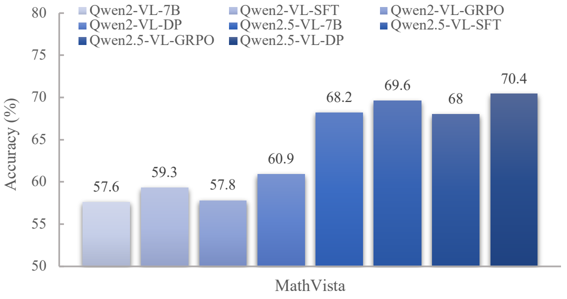


Figure 4: Accuracy of Qwen-VL model adopting different post-training strategies on MathVista.

5.3 Enhancements from SFT and RL

We conduct ablation study across three training paradigms: (1) supervised fine-tuning (SFT) on our curated dataset, (2) SFT followed by GRPO, and (3) RL applied in isolation. As shown in Figure 4, MLLM by SFT demonstrates improvements on the MathVista. Applying RL to the SFT model yields further gains, suggesting that RL facilitates deeper and more varied deductive reasoning. These progressive enhancements underscore the complementary strengths of SFT and RL: while SFT provides a stable foundation by aligning the model with diverse high-quality reasoning perspectives, RL further strengthens these abilities by promoting advanced cognitive behaviors. In contrast, applying RL without prior SFT leads to suboptimal performance, likely due to the absence of a structured reasoning baseline. Overall, integrating SFT with RL emerges as an effective paradigm for enhancing the MLLM's mathematical reasoning ability.

6 Conclusions

In this work, we proposed MathV-DP, a novel dataset that enriched multimodal mathematical reasoning with diverse solving perspectives and reflective supervision. Building upon Qwen-VL , we introduced Qwen-VL-DP, trained via both SFT and group relative policy optimization (GRPO), a rule-based reinforcement learning method tailored to reward correctness, diversity, and discrimination of multiple solutions. Our experiments on MathVista and Math-V benchmarks demonstrated that incorporating diverse reasoning perspectives significantly enhanced both the accuracy and generative diversity of MLLMs. These findings highlight the importance of moving beyond one-to-one image-text supervision, advocating for a shift towards learning from multiple valid solving perspectives.

7 Limitations

By learning from synthetic CoT data with diverse solving perspectives and reflection, and preference data involving discrimination of solution diversity and correctness, MLLM could be enhanced in multimodal mathematical reasoning as well as generative diversity across multiple responses. Such diversity could not be controlled explicitly in a single response; a single generation tends to randomly be one of the multiple correct solution perspectives learned. In future work, our model will be guided or trained to controllably generate the expected solution perspective.

8 Ethics Statement

We do not envision that our work will result in any harm as defined in ethics policy. Qwen2-VL and Qwen2.5-VL base model use Apache License. For datasets, MultiMath-300K uses Apache License 2.0. The evaluation datasets use permissive Creative Commons Licenses. The intended use of these source datasets and evaluation datasets is to train and test the model's multimodal reasoning capability, which is consistent with our utilization of all these data. Our proposed MathV-DP dataset can improve the multimodal mathematical reasoning ability of the open-source Qwen-VL through training.

References

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. Claude-3 Model Card.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. CoRR, abs/2308.12966.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

Feng Chen and Yujian Feng. 2023. Chain-of-thought prompt distillation for multimodal named entity and multimodal relation extraction. arXiv preprint arXiv:2306.14122.

Jiaxing Chen, Yuxuan Liu, Dehu Li, Xiang An, Ziyong Feng, Yongle Zhao, and Yin Xie. 2024. Plug-and-play grounding of reasoning in multimodal large language models. arXiv preprint arXiv:2403.19322.

Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. 2025. 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.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in Neural Information Processing Systems, 36.

K Anders Ericsson and Herbert A Simon. 1980. Verbal reports as data. Psychological review, 87(3):215.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. 2023. G-llava: Solving geometric problem with multi-modal large language model. arXiv preprint arXiv:2312.11370.

Google. Gemini. https://gemini.google.com.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Tora: A tool-integrated reasoning agent for mathematical problem solving. CoRR, abs/2309.17452.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948.

Yanyang Guo, Fangkai Jiao, Zhiqi Shen, Liqiang Nie, and Mohan S. Kankanhalli. 2023. UNK-VQA: A dataset and A probe into multi-modal large models' abstention ability. CoRR, abs/2310.10942.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. arXiv preprint arXiv:2305.02301.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. 2023. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. CoRR, abs/2312.03052.

Albert Qiaochu Jiang, Sean Welleck, Jin Peng Zhou, Timothée Lacroix, Jiacheng Liu, Wenda Li, Mateja Jamnik, Guillaume Lample, and Yuhuai Wu. 2023. Draft, sketch, and prove: Guiding formal theorem provers with informal proofs. In The Eleventh International Conference on Learning Representations.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, pages 14369–14387.

Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. 2023a. Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 9114–9128.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. 2023b. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. arXiv preprint arXiv:2311.07575.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In Advances in Neural Information Processing Systems.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. arXiv preprint arXiv:2406.15126.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. CoRR, abs/2310.02255.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. CoRR, abs/2308.09583.

Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. Reft: Reasoning with reinforced fine-tuning. arXiv preprint arXiv:2401.08967, 3.

OpenAI-a. Chatgpt. `https://chat.openai.com`.

OpenAI-b. Gpt-4v(ision). `https://openai.com/research/gpt-4v-system-card`.

OpenAI-c. Introducing openai o1. `https://openai.com/o1/`.

OpenBMB. 2024. Large multi-modal models for strong performance and efficient deployment. `https://github.com/OpenBMB/OmniLMM`.

Shuai Peng, Di Fu, Liangcai Gao, Xiuqin Zhong, Hongguang Fu, and Zhi Tang. 2024. Multimath: Bridging visual and mathematical reasoning for large language models. arXiv preprint arXiv:2409.00147.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36:53728–53741.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

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

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. 2024. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. arXiv preprint arXiv:2406.17294.

Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani. 2025. Evaluating the diversity and quality of llm generated content. arXiv preprint arXiv:2504.12522.

Richard S Sutton, Andrew G Barto, et al. 1998. Reinforcement learning: An introduction, volume 1. MIT press Cambridge.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. arXiv preprint arXiv:2501.12599.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. arXiv preprint arXiv:2402.14804.

Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2024b. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 19162–19170.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024c. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In The Eleventh International Conference on Learning Representations.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems.

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

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

Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad A. Ayyubi, Kai-Wei Chang, and Shih-Fu Chang. 2023. Idealgpt: Iteratively decomposing vision and language reasoning via large language models. In Findings of the Association for Computational Linguistics: EMNLP, pages 11289–11303.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023a. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. CoRR, abs/2311.16502.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023b. Mammoth: Building math generalist models through hybrid instruction tuning. CoRR, abs/2309.05653.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024. Rest-mcts*: Llm self-training via process reward guided tree search. Advances in Neural Information Processing Systems, 37:64735–64772.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. Advances in Neural Information Processing Systems, 36:5168–5191.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In The Eleventh International Conference on Learning Representations.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. CoRR, abs/2304.10592.

Wenwen Zhuang, Xin Huang, Xiantao Zhang, and Jin Zeng. 2025. Math-puma: Progressive upward multimodal alignment to enhance mathematical reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pages 26183–26191.

A Appendix

A.1 Evaluation on Generated Dataset

Human evaluation is a widely adopted method for assessing the quality of synthesized data (Long et al., 2024). We conduct a manual review of 1,000 randomly selected samples by five annotators to ensure the objectivity of checks. Like students marking exam papers, our evaluation emphasizes key aspects including overall correctness of solution and distinctions between correct and incorrect outputs. The average scores across these dimensions were 0.82 (±0.8%), 0.97 (±0.4%), and 0.95 (±0.5%) (on a scale of 0, 1), indicating that the generated solutions are generally of high quality and it is sufficiently reliable to enrich the solution space by introducing greater diversity. The synthesized solutions demonstrate varied mathematical reasoning approaches, offering a broader set of reasoning patterns that enhance the base model's capabilities. Additionally, we employ GPT-4o (OpenAI-a) to evaluate the generated solutions based on the original images and questions, filtering out duplicates and instances with inconsistent correctness labels. This process results in 38K examples for post-training on Qwen2.5-VL-7B but achieving an accuracy of 70.3% on MathVista, which is comparable to the performance of Qwen2.5-VL-DP. These results demonstrate both the quality assurance of our generated data and its effectiveness in enhancing MLLM's ability.

A.2 Generalizability of Qwen-VL-DP

The proposed Qwen-VL-DP model exhibits strong performance in multimodal mathematical reasoning tasks. To further evaluate its generalization capabilities, we conduct experiments on the MMMU benchmark (Yue et al., 2023a), which spans a wide range of disciplines and domains. As shown in Table 4, the Qwen-VL-DP model, post-trained on MathV-DP, consistently outperforms the base model as well as several other open-source MLLMs. These results highlight the model's ability to generalize effectively to diverse downstream multimodal understanding and reasoning tasks. Notably, the post-training with our synthetic dataset not only preserves but also enhances the model's reasoning performance in other domains, demonstrating the robustness and generalizability of Qwen-VL-DP.

Model	MMMU
Random Chance	22.1
Frequent Guess	26.8
SPHINX-13B	32.9
InstructBLIP-7B	32.9
LLaVA-1.5-13B	36.4
Qwen2-VL-7B	47.8
Qwen2-VL-DP	49.4
Qwen2.5-VL-7B	58.6
Qwen2.5-VL-DP	**59.4**

Table 4: Comparison on the MMMU benchmark.

A.3 Effectiveness of Disc./Preference Reward

To assess the individual impacts of the discrimination and preference reward components in our GRPO training, we conduct ablation experiments on Qwen2.5-VL-7B by selectively removing each reward. Specifically, we evaluate three variants: (1) removing both discrimination and preference rewards, (2) retaining the discrimination reward, and (3) retaining the preference reward. As shown in Table 5, the corresponding accuracies on the MathVista benchmark are respectively lower than the Qwen-VL-DP model. These results indicate that both rewards contribute positively to model performance, and their combined effect yields the better reasoning ability. The discrimination reward encourages the model to recognize diverse but valid solution paths, while the preference reward guides the model toward favoring correct over incorrect solutions. Together, they promote both reasoning diversity and solution correctness. The observed performance drop when either component is removed confirms their complementary roles and highlights the importance of explicitly rewarding diversity and correctness in the reasoning process.

Reward	w/ Both	w/o Disc.	w/o Pref.	w/o Both
MathVista	70.4	70.1	70.2	69.9

Table 5: Performance comparison by isolating discrimination and preference rewards.