

MM-Verify: Enhancing Multimodal Reasoning with Chain-of-Thought Verification

Lin Zhuang Sun^{1,5*}, Hao Liang^{2,4*}, Jingxuan Wei^{1,5†}, Bihui Yu^{1,5}, Tianpeng Li³,
Fan Yang^{3†}, Zenan Zhou^{3†}, Wentao Zhang^{2,4†}

¹University of Chinese Academy of Sciences

²Peking University

³Baichuan Inc. ⁴Zhongguancun Academy

⁵Shenyang Institute of Computing Technology, Chinese Academy of Sciences

sunlinzhuang21@mails.ucas.ac.cn, hao.liang@stu.pku.edu.cn

{yangfan, zhouzenan}@baichuan-inc.com, wentao.zhang@pku.edu.cn

Abstract

According to the Test-Time Scaling, the integration of External Slow-Thinking with the Verify mechanism has been demonstrated to enhance multi-round reasoning in large language models (LLMs). However, in the multimodal (MM) domain, there is still a lack of a strong MM-Verifier. In this paper, we introduce MM-Verifier and MM-Reasoner to enhance multimodal reasoning through longer inference and more robust verification. First, we propose a two-step MM verification data synthesis method, which combines a simulation-based tree search with verification and uses rejection sampling to generate high-quality Chain-of-Thought (COT) data. This data is then used to fine-tune the verification model, MM-Verifier. Additionally, we present a more efficient method for synthesizing MMCOT data, bridging the gap between text-based and multimodal reasoning. The synthesized data is used to fine-tune MM-Reasoner. Our MM-Verifier outperforms all larger models on the MathCheck, MathVista, and MathVerse benchmarks. Moreover, MM-Reasoner demonstrates strong effectiveness and scalability, with performance improving as data size increases. Finally, our approach achieves strong performance when combining MM-Reasoner and MM-Verifier, reaching an accuracy of 65.3 on MathVista, surpassing GPT-4o (63.8) with 12 rollouts. Our code is made available <https://github.com/Aurora-slz/MM-Verify>.

1 Introduction

Large language models (LLMs) have demonstrated exceptional performance across diverse tasks spanning myriad domains (OpenAI, 2023a; Touvron et al., 2023). Based on LLMs, MLLMs (Zhao et al., 2023; Wu et al., 2023; Bai et al., 2024) also show strong understanding ability among different modalities (Liu et al., 2023b; Bai et al.,

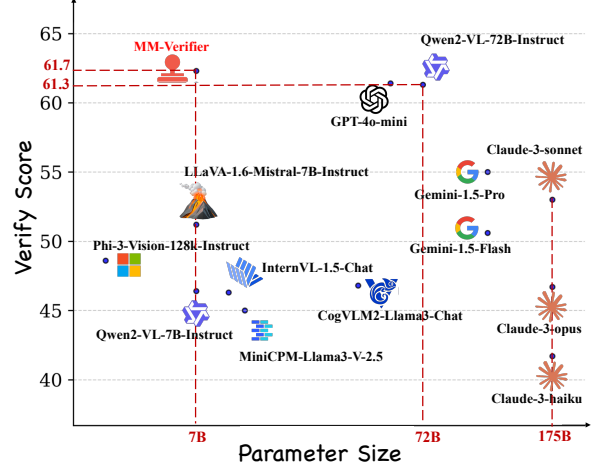


Figure 1: Our 7B MM-Verifier outperform all other models, even large models like GPT-4o, Gemini and Claude on the MathCheck Outcome-Judging benchmark.

2023b). They have demonstrated strong performance in image classification (Chen et al., 2024b), image understanding (Li et al., 2023b,c), image captioning (Bai et al., 2023b), visual question answering (Liu et al., 2023b,a) and image-text retrieval (Chen et al., 2024b). Recently, MLLMs have also made significant strides in solving mathematical problems (Liang et al., 2023; Huang et al., 2024). Researcher in our community have made efforts in designing strong reasoning models and algorithms (Thawakar et al., 2025; Du et al., 2025). Despite the effort made in processing MLLMs, we still face two challenges:

Lack of Strong MM Verifiers In pure-text LLMs, models can improve through self-critic methods (Weng et al., 2022; Sun et al., 2024). However, as shown in Table 6, such methods face challenges in enhancing performance in multimodal models. Therefore, it is crucial to develop robust multimodal (MM) verifiers.

Lack of Long COT Reasoning Data In the pure-text domain, models like DeepSeek-R1 (Guo

*Equal Contribution.

†Corresponding Author

et al., 2025), s1 (Muennighoff et al., 2025), and LIMO (Ye et al., 2025) have demonstrated the effectiveness of Long COT data. However, in the multimodal domain, most collected mathematics problems are not in the Long COT format (Li et al., 2024a; Lu et al., 2022; Chen et al., 2024a; Zhang et al., 2023; Shao et al., 2024; Cheng et al., 2025). Therefore, the development of Long COT synthetic methods is necessary to enhance the reasoning ability of MLLMs.

To address these issues, in this paper, we introduce two novel data synthesis methods and subsequently train MM-Verifier and MM-Reasoner. First, we perform a tree search, using simulations as rewards, to generate high-quality long MMCOT data. Next, we fine-tune MLLMs on our data, resulting in long chain-of-thought responses. These data are then Given that MLLMs generate long COT responses, we then apply rejection sampling to further enhance their verification capabilities, leading to the proposal of MM-Verifier. After the introduction of MM-Verifier, we observed that long COT data significantly improves model reasoning performance. As a result, we aim to synthesize large amounts of long COT data to enhance the performance of base MLLMs. Since tree search can be computationally expensive, we use the MAVIS dataset (Zhang et al., 2024e), leveraging the descriptions of patterns in MAVIS and inputting them into the pure-text reasoning model, Qwen QWQ. We then pair these patterns with the corresponding responses from the QWQ model. This method has proven to be effective, scalable, and capable of efficiently constructing large amounts of long MMCOT data.

The core contributions are summarized as follows:

- **MM Reasoning Data Synthesis Method** We propose two novel data synthesis methods for both our MM-Verifier and MM-Reasoner. First, we introduce a two-step MM-verification data synthesis approach that combines simulation-based tree search with GPT-4 verification and rejection sampling to generate high-quality COT data. Additionally, we use graphical software to link multimodal geometric shapes with textual descriptions, enabling the generation of multimodal reasoning data through a purely text-based reasoning model.
- **MM-Verifier** We introduce a new multimodal Outcome Reward Model (ORM) called MM-

Verifier. MM-Verifier achieves state-of-the-art (SOTA) performance on the MathCheck benchmark, surpassing closed-source models such as GPT-4, Gemini, and Claude. Furthermore, our MM-Verifier-7B outperforms Qwen2-VL-72B across all metrics on the MathVista and MathVerse benchmarks.

- **Scalability of MM-Reasoner** We propose a novel MM-Reasoning model based exclusively on our synthetic COT data. Although our MM-Reasoner does not achieve SOTA performance, it outperforms the baseline models and demonstrates scalability as the size of the training dataset increases. This provides new insights for the development of more powerful MM-Reasoners.
- **Strong Performance** By combining MM-Verifier and MM-Reasoning, with a model size of only 7B parameters, we outperform both GPT-4 and human performance on the MathVista benchmark, highlighting the strong performance of our method.

2 Related Work

2.1 MLLMs for Mathematics

Commonly Used MLLMs The integration of visual knowledge into LLMs has become a pivotal area of research due to the rapid advancements in LLMs. MLLMs combine vision information from vision encoders with LLMs, thus enabling these models to process and interpret visual inputs for various visual tasks (Liu et al., 2023c; Zhang et al., 2022; Li et al., 2022b) with enhanced accuracy and efficiency. Pioneering frameworks like CLIP (Radford et al., 2021) leverage contrastive learning on expansive image-caption datasets to align modalities, forming the groundwork for cross-modal comprehension. Various adapters (Liu et al., 2023b,a; Li et al., 2023b, 2022a; Jian et al., 2023; Lu et al., 2023a) are introduced to further integrate different modalities. For example, LLaVA (Liu et al., 2023b,a) employs a straightforward MLP to inject the vision information into LLMs. Whereas more complex implementations like the Q-Former in BLIP (Li et al., 2022a, 2023b) utilize cross-attention to enhance modality integration.

Recent studies (Wang et al., 2024b; Chen et al., 2023; Liu et al., 2023b,a; Li et al., 2023a; Zhang et al., 2024b; Zhuang et al., 2024; Luo

et al., 2025) aim to enhance MLLM performance by improving the quality of both pre-training and fine-tuning datasets. Models such as LLaVA (Liu et al., 2023b,a), ShareGPT4V (Chen et al., 2023), LLaVA-Next, LLaVA-OneVision (Liu et al., 2023b,a), Qwen2-VL, and Qwen2.5-VL (Bai et al., 2023b) have demonstrated significant advancements in understanding and executing complex instructions through instruction tuning. Leveraging large-scale training data, these models have also achieved strong performance in solving mathematical problems (Lu et al., 2023b).

MLLMs Designed for Math Problems In real-world applications, vision inputs are commonly used to present mathematical problems for models to solve. As a result, it is crucial for Vision-Language Large Models (MLLMs) to demonstrate strong mathematical capabilities. Meidani et al. (Meidani et al., 2023) pioneered the use of symbolic data to train a Vision-Language Model (VLM) with mathematical proficiency. Building on this work, UniMath (Liang et al., 2023) combined vision, table, and text encoders with LLMs, achieving state-of-the-art performance at the time. Additionally, Huang et al. (Huang et al., 2024) succeeded in solving algebraic problems that involved geometric diagrams.

Another noteworthy line of research involves using LLMs to tackle geometric problems. G-LLaVA (Gao et al., 2023) fine-tuned LLaVA (Liu et al., 2023b) with geometric data, reaching SOTA performance in geometry. Subsequently, MAVIS (Zhang et al., 2024e) and EAGLE (Li et al., 2024b) achieved SOTA results by introducing math-specific encoders and amassing large amounts of mathematical data.

2.2 LLM-as-a-Judge

In the Reinforcement Learning from Human Feedback (RLHF) or MCTS-based inference, Reward Models (RMs) are employed to assess and score the quality of model outputs, thereby guiding the optimization or reasoning path of LLMs (Chen et al., 2025). Reward models can be categorized into Process Reward Models (PRMs) and Outcome Reward Models (ORMs).

Outcome Reward Models. ORM evaluates only the final mathematical results without considering the solution process. For instance, Qwen2.5-Math-RM-72B (Zhang et al., 2025), released by the Qwen team, assigns a single score to each mathe-

matical response.

Process Reward Models. PRMs are more fine-grained, focusing on whether each step of the reasoning path is logical and correct, providing step-level feedback and guidance signals. For example, Math-Shepherd (Wang et al., 2024a) is trained on an automatically constructed (rather than manually annotated) process supervision dataset, scoring each step of mathematical reasoning. MATHMinos-PRM (Gao et al., 2024) introduces a novel two-stage training paradigm and incorporates step-wise natural language feedback labels. EurusPRM (Cui et al., 2025) utilizes implicit PRM, where ORM is trained to evaluate response-level labels. Qwen2.5-Math-PRM (Zhang et al., 2025), currently the SOTA PRM, proposes a consensus filtering mechanism combining Monte Carlo estimation and LLM-as-a-judge. Additionally, there are the Skywork-PRM series (o1 Team, 2024) and RLHFlow-PRM series (Xiong et al., 2024) models. Moreover, Liu et al. (2024) proposed Multimodal PRM based on Monte Carlo rollouts. For more comprehensive LLM-as-a-Judge please refer to the LLM-as-a-Judge survey (Gu et al., 2024).

3 Methodology

Table 1: Statistical details of our collected data.

Dataset	Subdataset	Number	Ratio
MM-Verify	Geometry3K	20226	33.84%
	FigureQA	10800	18.07%
	GEOS	882	1.48%
	Super-CLEVR	14446	24.17%
	TabMWP	13418	22.45%
	sum	59772	100%
MM-Reasoner	MAVIS-Geo	32146	100%

In this section, we introduce the construction process of MM-Verifier, as illustrated in Figure 2. Section 3.1 details the data synthesis methodology for MM-Verifier (Stage 1). Section 3.2 describes the data synthesis scheme employed in MM-Verifier (Stage 2). In Section 3.3, we explore the enhancement of multimodal model reasoning capabilities through the integration of long-COT data in pure text form.

3.1 Stage1: Long CoT MM-Verifier

3.1.1 Source Data Collection

MM-Verifier is designed to verify the correctness of a $\langle q, s \rangle$ pair by determining whether the

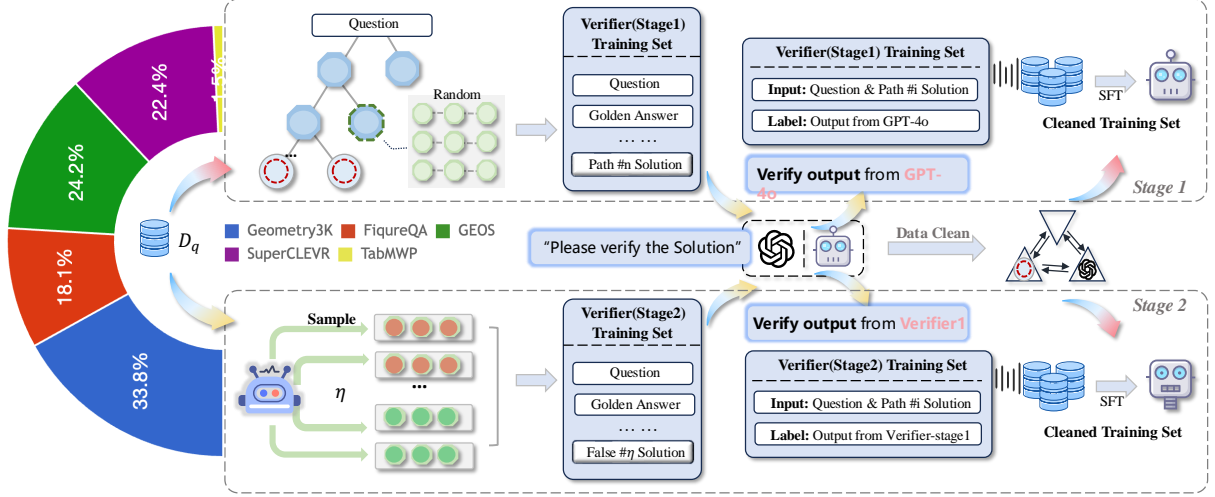


Figure 2: We present the pipeline for synthesizing MM-Verifier data. In Stage 1, we use a simulation-based algorithm for long-chain CoT reasoning and long verification. In Stage 2, we use the trained Verifier model from Stage 1 to further enhance it using rejection sampling, generating more long CoT verification data.

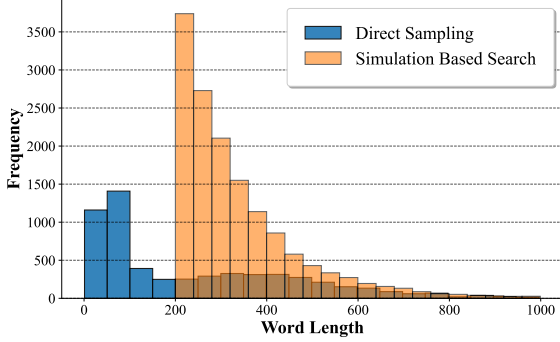


Figure 3: Answer length of direct sampling and simulated-based search. We can see the simulated-based search can synthesize longer COT answers.

solution s is correct. To achieve this goal, it is essential to synthesize diverse verification data from a variety of sources. Specifically, we construct the question source pool D_s from seven categories in MATH360V: Geometry3K, TabMWP, SuperCLEVR, UniGeo, FigureQA, and GEOS, with their statistical details summarized in Table 1. By collecting data from these categories, we obtained a diverse set of multimodal mathematics questions.

However, solutions in these datasets are typically very short, and many contain only answers. When verifying reasoning, the answer is often provided in a long CoT form. Therefore, we need to construct CoT data ranging from short to long, along with their corresponding verifications. To facilitate the generation of long CoT data and enhance the diversity of the training set, we design a simulation-based search algorithm to create extended reasoning trajectory data for training MM-Verifier.

3.1.2 Simulation-based Search Algorithm

Inspired by Monte Carlo Tree Search (MCTS), we propose a simulation-based search algorithm tailored for multimodal models. However, we do not directly apply the traditional MCTS tree search, as multimodal models often fail to generate reliable rewards, which results in suboptimal performance, as illustrated in Appendix A. To address this challenge, we introduce a simulation-based reward mechanism.

Starting from the root node q_i , we first simulate k child nodes for each node. For each child node, we perform simulations where the model directly generates answers based on the current node. For a tree node u_d , which represents a node at depth d , the ancestral path leading up to the root is denoted by the sequence $\{u_{d-1}, \dots, u_1\}$. The simulation answer for this path is given by:

$$\text{Simulation Answer} = LLM \left(\bigoplus_{i=1}^{d-1} u_i \right)$$

These simulations are repeated l times, and the correctness ratio of the *Simulation Answer* is used as the reward. Once the reward is obtained, we apply the MCTS algorithm for further simulation and data synthesis.

Using the simulation-based MCTS approach, for each question, we perform n rollouts and collect solution pairs $\langle q_i, p_j^i \rangle$, where $j \in \{1, 2, \dots, n\}$ and n represents the number of leaf nodes. These n solution pairs are then verified as positive and negative cases for training the MM-verifier.

Table 2: We compare performance of Qwen2-VL-Instruct-7B/72B, LLaMA-3.2-11B-Vision-Instruct with our MM-Reasoner on the MathVista testmini benchmark. GVQA: General VQA; MVQA: Math Target VQA.

Method	Qwen2-VL			LLaMA-3.2-11B-Vision			MM-Reasoner		
	ALL	GVQA	MVQA	ALL	GVQA	MVQA	ALL	GVQA	MVQA
<i>Sample 4</i>									
Majority Voting	57.1	65.7	49.8	45.2	50.8	40.4	59.4	65.2	54.3
Qwen2-VL-7B as Judgement	57.9	66.3	50.7	41.7	48.9	35.5	53.1	59.8	47.4
Qwen2-VL-72B as Judgement	53.4	60.4	47.4	46.3	49.8	43.3	53.7	59.8	48.5
MM-Verifier(Stage1)	58.8	64.1	54.3	50.0	56.3	44.6	60.6	66.5	53.7
MM-Verifier(Stage2)	59.8	67.0	53.7	50.4	55.9	45.7	61.5	65.2	58.3
<i>Sample 8</i>									
Majority Voting	61.1	68.5	54.8	48.3	52.8	44.4	62.2	68.0	57.2
Qwen2-VL-7B as Judgement	54.5	62.0	48.1	46.1	50.7	42.2	53.6	61.5	46.9
Qwen2-VL-72B as Judgement	56.2	62.4	50.9	46.2	51.1	42.0	53.9	61.1	47.8
MM-Verifier(Stage1)	61.6	66.5	57.4	51.4	56.3	47.2	63.8	68.7	59.6
MM-Verifier(Stage2)	62.5	68.5	57.4	52.1	57.6	47.4	65.3	69.8	61.5
<i>Sample 12</i>									
Majority Voting	62.9	68.5	58.1	51.3	56.5	46.9	64.8	68.7	61.5
Qwen2-VL-7B as Judgement	54.4	60.0	49.6	45.5	48.6	42.0	55.4	61.1	50.6
Qwen2-VL-72B as Judgement	55.7	61.1	51.1	46.5	49.8	43.7	55.6	61.5	48.7
MM-Verifier(Stage1)	63.7	70.2	58.1	55.0	59.8	50.9	64.3	69.1	60.1
MM-Verifier(Stage2)	64.1	70.4	58.7	55.9	60.9	51.7	65.2	69.8	61.3

3.1.3 Long COT Verification Data Synthesize

After obtaining the $\langle q_i, p_j^i \rangle$ pair, the next step is to verify p_j^i to determine whether q_i has been answered correctly. To do this, we use GPT-4o (gpt-4o-2024-08-06) to verify $\langle q_i, p_j^i \rangle$ using the instruction "Verify step by step..." (for the detailed prompt, see Figure 7). The model's output, denoted as v_i , serves as the target for the verifier. The resulting dataset collected at this stage is represented as D_v , which can be expressed as:

$$D_v = \{(q_i, p_j^i, v_i) \mid i = 1, \dots, m; j = 1, \dots, n\}$$

Additionally, we implement a data-cleaning strategy to filter high-quality synthetic data for training the MM-Verifier. For each data instance (q_i, p_j^i, v_i) in D_v , we first design an answer extraction prompt, as shown in Figure 9) and use LLaMA-3.2-3B-Instruct for answer extraction, denoted by $\text{extract}()$ for clarity. We then apply the following criteria for data cleaning:

- **Condition 1:** If $\text{extract}(p_j^i)$ matches the golden label $\text{extract}(y_i)$, and the final result of v_i is the answer is correct.
- **Condition 2:** If $\text{extract}(p_j^i)$ does not match $\text{extract}(y_i)$, and the final result of v_i is the answer is not correct.

We collect the instances that meet the above conditions, (q_i, p_j^i, v_i) , into D_{clean} . Any instance that

does not satisfy these conditions is discarded. The dataset D_{clean} is then used for supervised fine-tuning (SFT) on Qwen2-VL-7B-Instruct, yielding the first-stage verifier, MM-Verifier (Stage 1).

3.2 Stage 2: Rejection Sampling Further Improves Verification

In Stage 1, we obtained a Verifier with strong long-chain Chain-of-Thought (CoT) reasoning capabilities. In Stage 2, our goal is to improve the efficiency of the data synthesis process, reduce API costs, and further enhance the Verifier's capabilities.

To achieve this, we first generate corresponding solutions based on a given set of questions. By leveraging the long CoT reasoning ability of the MM-Verifier from Stage 1, we can generate additional long CoT verification data.

The synthetic data is then cleaned using string matching against the correct answer. The filtered data is subsequently fed into the MM-Verifier (Stage 1) for further training, resulting in the enhanced MM-Verifier (Stage 2).

3.3 Bridging the Gap Between Text and MM

When applying MM-Verifier, the base model generates multiple reasoning paths for a given question. The MM-Verifier then evaluates these paths, distinguishing between correct and incorrect inferences. This process inherently requires the base model to produce at least one plausible correct reasoning path for the MM-Verifier to recognize

Table 3: We compare performance of Qwen2-VL-Instruct-7B/72B, LLaMA-3.2-11B-Vision-Instruct with our MM-Reasoner on the MathVerse testmini benchmark. VD: Vision Dominant; VI: Vision Intensive; TL: Text Lite.

Method	Qwen2-VL				llama-3.2-11B-Vision				MM Reasoner			
	ALL	VD	VI	TL	ALL	VD	VI	TL	ALL	VD	VI	TL
<i>Sample 4</i>												
Majority Voting	20.1	19.8	19.0	17.7	17.8	15.9	17.4	19.8	22.9	20.7	23.0	23.2
Qwen2-VL-7B as Judgement	20.6	22.3	20.8	19.0	20.6	19.3	21.1	21.2	22.6	21.4	21.4	24.6
Qwen2-VL-72B as Judgement	21.5	20.7	21.1	22.0	20.2	20.8	18.1	21.6	23.0	22.0	22.3	24.0
MM-Verifier(Stage1)	24.0	25.3	22.6	22.8	21.9	23.2	20.8	22.5	24.8	22.2	25.6	26.0
MM-Verifier(Stage2)	24.3	23.6	23.2	23.1	22.4	21.8	23.0	24.4	25.3	23.1	23.5	27.4
<i>Sample 8</i>												
Majority Voting	22.9	21.8	22.1	23.5	23.5	20.6	22.8	25.5	24.5	20.6	25.0	24.5
Qwen2-VL-7B as Judgement	21.1	20.6	20.7	21.1	21.1	19.7	21.1	20.0	21.7	19.2	24.6	20.3
Qwen2-VL-72B as Judgement	21.5	20.3	21.2	21.2	20.8	19.5	20.2	20.4	22.3	19.4	24.4	23.1
MM-Verifier(Stage1)	25.1	22.7	24.2	25.5	24.8	21.6	24.4	25.4	25.3	23.1	23.5	27.4
MM-Verifier(Stage2)	25.2	24.2	24.1	25.1	25.0	22.7	26.3	25.9	25.7	23.2	25.3	26.3
<i>Sample 12</i>												
Majority Voting	24.8	19.9	24.7	25.6	24.0	19.7	24.5	26.1	25.9	23.1	26.1	27.3
Qwen2-VL-7B as Judgement	22.0	20.1	22.0	22.5	20.0	18.9	19.9	23.1	21.8	19.5	22.7	23.4
Qwen2-VL-72B as Judgement	22.3	20.4	22.3	23.0	20.7	19.4	20.9	20.7	22.1	20.1	23.2	25.0
MM-Verifier(Stage1)	25.7	23.1	26.5	26.5	24.5	21.1	23.2	25.5	27.0	23.5	27.0	29.4
MM-Verifier(Stage2)	25.8	21.7	25.4	28.2	24.7	21.2	24.7	27.0	27.3	24.1	27.7	28.9

and validate. Therefore, in addition to a promising MM-Verifier, In this section, our goal is to synthesize long COT data to train a robust MM-Reasoner capable of learning long COT reasoning. However, synthesizing long COT data using tree search can be computationally expensive. Moreover, long COT pure text models have demonstrated strong performance. To efficiently generate long COT data, we propose distilling knowledge from pure text models to our MM-Reasoner.

Specifically, we select data from MAVIS-GEOMETRY (Zhang et al., 2024e), which includes geometric pattern drawings along with textual instructions. By combining the geometric textual instructions with the original questions, we can feed them into a pure text reasoning model. The outputs generated by the reasoning model, Qwen-QwQ (Team, 2024) with prompts in Figure 8, are then collected as target labels for the MM-Reasoner training dataset, denoted as D_r . To ensure the quality of the data, we filter out incorrect QwQ-generated reasoning results from D_r . Then we use the filtered data to train Qwen2-VL-7B-Instruct with supervised fine-tuning (SFT), ultimately obtaining the MM-Reasoner model.

4 Experiments

4.1 Experiment Setting

Baseline Models. The base models for MM-Verifier and MM-Reasoner are Qwen2-VL-Instruct-7B (Bai et al., 2023b). For comparison, we include random selection and human performance as baselines, along with two types of closed-source models: GPT-4o (gpt-4o-2024-08-06) (OpenAI, 2023b) and Qwen-VL-Plus (Bai et al., 2023a). For open-source models, we evaluate ten MLLMs: mPLUG-Owl2-7B (Ye et al., 2024), MiniGPT4-7B (Zhu et al., 2023), LLaVA-1.5-13B (Liu et al., 2023a), SPHINX-V2-13B (Lin et al., 2023), Deepseek-VL (Lu et al., 2024), LLaVA-OneVision-7B (llava-onevision-qwen2-7b-ov-hf) (Li et al., 2024a), Qwen2-VL-Instruct-7B (Bai et al., 2023b), Llama-3.2-11B-Vision (Touvron et al., 2023), Math-LLaVA (Shi et al., 2024), and G-LLaVA-7B (Gao et al., 2023).

MM-Verifier Baselines. **Majority Voting:** Select the answer that appears most frequently among multiple candidate answers. **Qwen2-VL-7B and Qwen-VL-72B as Judgment:** We use Qwen2-VL-7B-Instruct and Qwen-VL-72B (Bai et al., 2023b) as the judgment model to assess the correctness of each candidate solution. If multiple candidates are deemed correct, we apply a majority voting

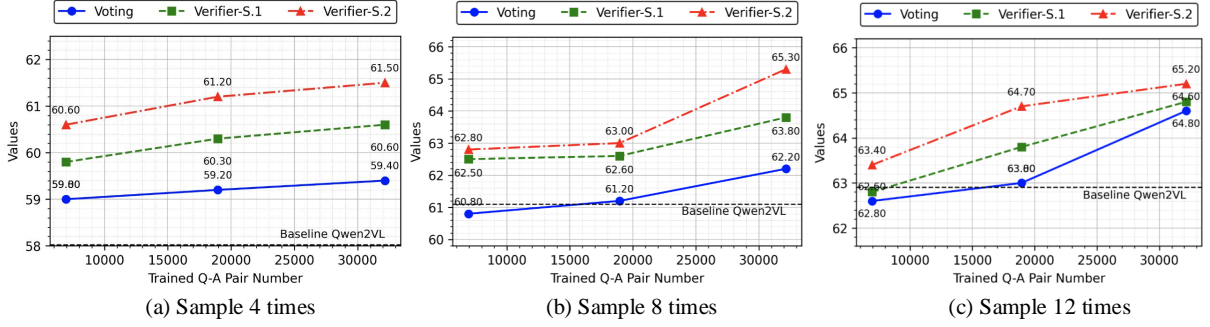


Figure 4: The performance of our MM-Reasoner can scale up using different MM-Verifiers. We can see with different scale MM-Reasoner the MM-Verifier consistently outperform majority voting and MM-Verifier Stage1.

mechanism to determine the final answer.

Benchmarks. We evaluate the performance of MM-Verifier, MM-Reasoner, and the baselines on two commonly used benchmarks: MathVista (Lu et al., 2023b) and MathVerse (Zhang et al., 2024d). Additionally, we assess the MM-Verifier on the Verify bench MathCheck (Multimodal Outcome Judging) (Zhou et al., 2024).

Settings. The settings include a maximum token limit of 4096, a top-k value of 5, a temperature of 0.3, and a repetition penalty of 1.05. All experiments are conducted on 8 NVIDIA H20 GPUs.

4.2 Effectiveness of MM-Verifier and MM-Reasoner

In this section, we demonstrate the effectiveness of both the MM-Verifier and MM-Reasoner. As shown in Figure 1, the MM-Verifier outperforms all other models in the MathCheck benchmark. Moreover, Tables 2 and 3 reveal that the MM-Verifier achieves strong performance on the MathVista and MathVerse benchmarks. Our results indicate that the MM-Verifier surpasses both the Majority Voting and Qwen2-VL-72B-Instruct methods. Specifically, MM-Verifier (Stage 2) delivers superior results, demonstrating its robust ability to verify answers and enhance model performance.

Furthermore, the MM-Reasoner outperforms powerful models, such as Qwen2-VL-Instruct-7B and LLaMA-3.2-11B-Vision, across all evaluation metrics. These findings clearly demonstrate that both the MM-Verifier and MM-Reasoner contribute significantly to performance improvements, underscoring their potential for addressing complex multimodal reasoning and verification tasks.

Interestingly, we observe that the performance of Qwen2-VL-72B, when used as a Judgment model, improves when verifying answers for Qwen2-VL

and LLaMA-3.2-11B-Vision (from sample 4 to sample 12). However, its performance drops when verifying MM-Reasoner outputs. This discrepancy arises because conventional models struggle to verify the correctness of longer outputs, while our MM-Verifier consistently maintains robust performance. This further emphasizes the versatility of the MM-Verifier across a wide range of scenarios.

4.3 Scalability of Our MM-Reasoner

The results in Figure 4 illustrate the scalability of MM-Reasoner with respect to the quantity of training data. As the amount of training data increases, the performance of the model consistently improves across all evaluation settings. Specifically, MM-Reasoner achieves steady performance gains when moving from 6,952 to 32,146 training samples, demonstrating its ability to effectively utilize larger datasets for better reasoning.

Additionally, both Verifier-S.1 and Verifier-S.2 show clear improvements as the training data grows, with Verifier-S.2 outperforming Verifier-S.1 in all cases. This trend highlights the effectiveness of the staged verification approach in enhancing reasoning accuracy.

These results emphasize the superiority of our method, as MM-Reasoner scales effectively with increasing training data, achieving higher performance and showcasing the robustness and adaptability of our approach.

4.4 MM-Verifier and MM-Reasoner achieved SOTA Performance

We leveraged our proposed MM-Reasoner and MM-Verifier together to enhance multimodal mathematical reasoning. Specifically, MM-Reasoner generated 12 diverse reasoning rollouts per query, while MM-Verifier systematically evaluated and filtered these rollouts, ensuring high-quality and

Table 4: We compare our MM-Verifier plus MM-Reasoner with closed and open source MLLMs, and other baselines. GPS: geometry problem solving; MWP: math word problem; FQA: figure question answering; TQA: textbook question answering; VQA: visual question answering.

Model	MATHVISTA						MATHVERSE				
	ALL	GPS	MWP	FQA	TQA	VQA	ALL	TD	TL	VI	VD
<i>Closed Source MLLMs & Other Baselines</i>											
Random	17.9	21.6	3.8	18.2	19.6	26.3	12.4	12.4	12.4	12.4	12.4
Human	60.3	48.4	73.0	59.7	63.2	55.9	64.9	71.2	70.9	61.4	68.3
GPT-4o	63.8	64.7	-	-	-	-	50.8	59.8	50.3	48.0	46.5
Qwen-VL-Plus	43.3	35.5	31.2	54.6	48.1	51.4	21.3	26.0	21.2	18.5	19.1
<i>Open-Source MLLMs</i>											
mPLUG-Owl2-7B	22.2	23.6	10.2	22.7	27.2	27.9	10.3	11.6	11.4	11.1	9.4
MiniGPT4-7B	23.1	26.0	13.4	18.6	30.4	30.2	12.2	12.3	12.9	12.5	14.8
LLaVA-1.5-13B	27.7	22.7	18.9	23.8	43.0	30.2	12.7	17.1	12.0	12.6	12.7
SPHINX-V2-13B	36.7	16.4	23.1	54.6	41.8	43.0	16.1	20.8	14.1	35.2	28.9
Deepseek-VL	34.9	28.4	55.9	26.8	32.9	34.6	19.3	23.0	23.2	20.2	18.4
LLaVA-OneVision-7B	51.1	66.8	54.8	39.4	57.6	40.8	20.7	25.5	21.8	20.9	21.2
Qwen2-VL-7B	52.5	44.7	59.7	55.4	52.5	49.7	20.1	23.7	16.3	19.0	19.8
llama-3.2-11B-Vision	46.5	38.0	51.6	50.9	51.9	39.7	17.8	19.2	17.9	15.6	15.5
G-LLaVA-7B	25.1	48.7	3.6	19.1	25.0	28.7	16.6	20.9	20.7	17.2	14.6
Math-LLaVA-13B	46.6	57.7	56.5	37.2	51.3	33.5	22.9	27.3	24.9	24.5	21.7
Ours	65.3	60.1	73.1	68.8	67.7	55.9	25.7	30.2	26.3	25.3	23.2

Input <question, solution>	
Q: As shown in the figure, the diameter CD of $\odot O$ crosses the midpoint G of chord EF, $\angle DCF = 20.0$, then $\angle EOD$ is equal to ()	
A: Let's think step by step.	
Given that CD is the diameter of circle O, and G is the midpoint of chord EF, thus OG is perpendicular to EF. Since $\angle DCF = 20^\circ$, and CD is the diameter, $\angle CED = 90^\circ$. Therefore, $\angle CGF = 90^\circ - 20^\circ = 70^\circ$. Since G is the midpoint of EF, $\angle CGF = \angle EGF = 70^\circ$. In triangle EOG, since G is the midpoint of EF, $\angle EOD = 2\angle EGF = 2 \times 70^\circ = 140^\circ$. Therefore, the answer is (D). The final answer is 10° .	
Verify output from MM-Verifier	
Let's verify the solution step by step:	
Understanding the Problem:	
1. Given: $\angle DCF = 20^\circ$, and G is the midpoint of EF.	
2. We need to find $\angle EOD$.	
Solution Steps Verification:	
1. Step 1: Since G is the midpoint of EF, triangle OEF is isosceles with $OE = OF$.	
2. Step 2: In circle O, $\angle DCF = 20^\circ$, which means $\angle DOE$ is a central angle subtending the same arc as $\angle DCF$.	
3. Step 3: Therefore, $\angle DOE = 2 \times \angle DCF = 2 \times 20^\circ = 40^\circ$.	
Conclusion:	
1. The solution correctly finds $\angle DOE = 40^\circ$.	
2. Since $\angle EOD$ is the same angle as $\angle DOE$, it should also be 40° .	
3. The final answer given is 10° , which is incorrect.	
Verification: Is the answer correct (Yes/No)? No.	✓
Verify output from Qwen2-VL-72B-Instruct	
Yes	✗

Figure 5: We present a case of MM-Verifier. We can see MM-Verifier correctly verify the answer with Long COT while Qwen2-VL-72B-Instruct failed to.

accurate responses. This iterative verification-refinement process significantly improved reasoning precision and robustness.

As shown in Table 4, our approach outperforms both open-source and closed-source MLLMs across multiple benchmarks. On the MATHVISTA dataset, our method achieves an overall accuracy of **65.3**, surpassing human performance

(**60.3**) and even GPT-4o (**63.8**). Similarly, in the MATHVERSE benchmark, our method consistently achieves strong performance with an overall score of **25.7**, outperforming strong baselines like Math-LLaVA-13B (**22.9**) and LLaVA-OneVision (**20.7**).

These results demonstrate the robustness of MM-Reasoner and MM-Verifier in improving mathematical reasoning across diverse tasks.

4.5 Case Study

Figure 5 presents a case study of MM-Verifier, demonstrating its ability to successfully identify logical errors in the reasoning process through step-by-step verification. In contrast, Qwen2-VL-72B-Instruct fails to provide a step-by-step reasoning trajectory, leading to an error in detecting these mistakes. This case underscores the superior analytical capabilities of MM-Verifier.

5 Conclusion

In this paper, we propose two data synthesis methods: the first generates long COT verification data, while the second synthesizes long COT inference data more efficiently. We use these synthetic data to train the MM-Verifier and MM-Reasoner. Our MM-Verifier not only outperforms larger models on the MathCheck benchmark but also demonstrates superior performance against larger models on benchmarks such as MathVista and MathVerse. Addi-

tionally, the MM-Reasoner exhibits strong scalability, with performance improving as the data size increases. Furthermore, the combination of MM-Verifier and MM-Reasoner achieves impressive results on the MathVista benchmark, surpassing even GPT-4o. These findings confirm the effectiveness of MM-Verifier and MM-Reasoner in enhancing multimodal reasoning tasks and lay the foundation for future advancements in this domain.

6 Limitations

Due to limited funding and computational resources, we were unable to scale our MM-Verifier and MM-Reasoner to Qwen2-VL-72B. Additionally, our scalability tests were restricted to datasets of fewer than 100K samples. We plan to conduct further experiments as soon as additional computational resources become available.

7 Acknowledgements

This work is supported by the National Key R&D Program of China (2024YFA1014003), National Natural Science Foundation of China (92470121, 62402016), CAAI-Ant Group Research Fund, and High-performance Computing Platform of Peking University.

References

- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.
- Tianyi Bai, Hao Liang, Binwang Wan, Ling Yang, Bozhou Li, Yifan Wang, Bin Cui, Conghui He, Binhang Yuan, and Wentao Zhang. 2024. A survey of multimodal large language model from a data-centric perspective. *arXiv preprint arXiv:2405.16640*.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multi-modal models with better captions. *CoRR*, abs/2311.12793.
- Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*.
- Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024a. M 3 cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv preprint arXiv:2405.16473*.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024b. 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*, pages 24185–24198.
- Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and Libo Qin. 2025. Comt: A novel benchmark for chain of multi-modal thought on large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23678–23686.
- Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, et al. 2025. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*.
- Yifan Du, Zikang Liu, Yifan Li, Wayne Xin Zhao, Yuqi Huo, Bingning Wang, Weipeng Chen, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. 2025. Virgo: A preliminary exploration on reproducing o1-like mllm. *arXiv preprint arXiv:2501.01904*.
- Bofei Gao, Zefan Cai, Runxin Xu, Peiyi Wang, Ce Zheng, Runji Lin, Keming Lu, Junyang Lin, Chang Zhou, Wen Xiao, et al. 2024. Llm critics help catch bugs in mathematics: Towards a better mathematical verifier with natural language feedback. *CoRR*.
- Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wan-jun 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*.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.
- 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*.
- Litian Huang, Xinguo Yu, Feng Xiong, Bin He, Shengbing Tang, and Jiawen Fu. 2024. Hologram reasoning for solving algebra problems with geometry diagrams. *arXiv preprint arXiv:2408.10592*.
- Yiren Jian, Chongyang Gao, and Soroush Vosoughi. 2023. Bootstrapping vision-language learning with decoupled language pre-training. In *Advances in Neural Information Processing Systems 36: Annual*

- Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.*
- Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *CoRR*, abs/2305.03726.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023c. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022a. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162, pages 12888–12900.
- Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. 2022b. Grounded language-image pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10955–10965. IEEE.
- Zhihao Li, Yao Du, Yang Liu, Yan Zhang, Yufang Liu, Mengdi Zhang, and Xunliang Cai. 2024b. Eagle: Elevating geometric reasoning through llm-empowered visual instruction tuning. *arXiv preprint arXiv:2408.11397*.
- Zhenwen Liang, Tianyu Yang, Jipeng Zhang, and Xiangliang Zhang. 2023. Unimath: A foundational and multimodal mathematical reasoner. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7126–7133.
- Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. 2023. 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. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2023c. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. *CoRR*, abs/2303.05499.
- Wei Liu, Junlong Li, Xiwen Zhang, Fan Zhou, Yu Cheng, and Junxian He. 2024. Diving into self-evolving training for multimodal reasoning. *arXiv preprint arXiv:2412.17451*.
- Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.
- Junyu Lu, Ruyi Gan, Dixiang Zhang, Xiaojun Wu, Ziwei Wu, Renliang Sun, Jiaying Zhang, Pingjian Zhang, and Yan Song. 2023a. Lyrics: Boosting fine-grained language-vision alignment and comprehension via semantic-aware visual objects. *CoRR*, abs/2312.05278.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023b. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.
- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Øyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.
- Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. 2025. Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. *arXiv preprint arXiv:2501.04686*.
- Kazem Meidani, Parshin Shojaei, Chandan K Reddy, and Amir Barati Farimani. 2023. Snip: Bridging mathematical symbolic and numeric realms with unified pre-training. *arXiv preprint arXiv:2310.02227*.
- Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, et al. 2024. Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems. *arXiv preprint arXiv:2412.09413*.
- Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke

- Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.
- Skywork o1 Team. 2024. *Skywork-o1 open series*. <https://huggingface.co/Skywork>.
- OpenAI. 2023a. *Chatgpt*.
- R OpenAI. 2023b. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5).
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642.
- 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*.
- Linzhuang Sun, Hao Liang, Jingxuan Wei, Bihui Yu, Conghui He, Zenan Zhou, and Wentao Zhang. 2024. Beats: Optimizing llm mathematical capabilities with backverify and adaptive disambiguate based efficient tree search. *arXiv preprint arXiv:2409.17972*.
- Qwen Team. 2024. *Qwq: Reflect deeply on the boundaries of the unknown*.
- Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. 2025. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024a. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439.
- Weizhi Wang, Khalil Mrini, Linjie Yang, Sateesh Kumar, Yu Tian, Xifeng Yan, and Heng Wang. 2024b. Finetuned multimodal language models are high-quality image-text data filters. *CoRR*, abs/2403.02677.
- Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2022. Large language models are better reasoners with self-verification. *arXiv preprint arXiv:2212.09561*.
- Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. 2023. Multimodal large language models: A survey. *arXiv preprint arXiv:2311.13165*.
- Wei Xiong, Hanning Zhang, Nan Jiang, and Tong Zhang. 2024. *An implementation of generative prm*. <https://github.com/RLHFlow/RLHF-Reward-Modeling>.
- Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multimodal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051.
- Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*.
- Di Zhang, Xiaoshui Huang, Dongzhan Zhou, Yuqiang Li, and Wanli Ouyang. 2024a. Accessing gpt-4 level mathematical olympiad solutions via monte carlo tree self-refine with llama-3 8b. *arXiv preprint arXiv:2406.07394*.
- Di Zhang, Jingdi Lei, Junxian Li, Xunzhi Wang, Yujie Liu, Zonglin Yang, Jiatong Li, Weida Wang, Suorong Yang, Jianbo Wu, et al. 2024b. Critic-v: Vlm critics help catch vlm errors in multimodal reasoning. *arXiv preprint arXiv:2411.18203*.
- Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. 2022. Glipv2: Unifying localization and vision-language understanding. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2024c. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*.
- Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. 2024d. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*.

- Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. 2024e. Mavis: Mathematical visual instruction tuning. *arXiv preprint arXiv:2407.08739*.
- Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301*.
- Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multi-modal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.
- Zihao Zhou, Shudong Liu, Maizhen Ning, Wei Liu, Jindong Wang, Derek F Wong, Xiaowei Huang, Qifeng Wang, and Kaizhu Huang. 2024. Is your model really a good math reasoner? evaluating mathematical reasoning with checklist. *arXiv preprint arXiv:2407.08733*.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.
- Wenwen Zhuang, Xin Huang, Xiantao Zhang, and Jin Zeng. 2024. Math-puma: Progressive upward multi-modal alignment to enhance mathematical reasoning. *arXiv preprint arXiv:2408.08640*.

A Performance of Naive MCTS in Multimodal Reasoning

In NLP-based mathematical reasoning tasks, given a question, the Monte Carlo Tree Search (MCTS) algorithm iteratively refines its responses through continuous self-reflection, ultimately converging to an optimized answer (Zhang et al., 2024a). Experimental results demonstrate that this search-based approach yields strong performance. Naturally, we hypothesized that applying the same methodology to multimodal tasks would lead to similar improvements. However, as shown in Table 6, we compare the performance of MCTS with Majority Voting and our MM-Verifier on the MathVista benchmark. We can see simple MCTS failed to improve reasoning performance. We attribute this to the significantly higher prevalence of hallucinations in multimodal models compared to their text-only counterparts. Therefore inspired by (Wang et al., 2024a; Sun et al., 2024) we design a simulation-based tree search.

B Detailed Performance of MM-Verifier on Sub-tasks of MathVista

Table 5 presents a comparative evaluation of Qwen2-VL-Instruct-7B, LLaMA-3.2-11B-Vision, and our proposed MM-Reasoner on the MathVista benchmark. The models were assessed across multiple reasoning-intensive sub-tasks, including GPS, MWP, FQA, TQA, and VQA.

In the Sample 4 evaluation, MM-Reasoner with the MM-Verifier (Stage 2) achieved an overall accuracy of 61.5, surpassing Qwen2-VL’s 59.8 and LLaMA-3.2-11B-Vision’s 50.4. This trend persists across Sample 8 and Sample 12, where MM-Reasoner obtained 65.3 and 65.7 respectively, further establishing its robustness in complex multimodal reasoning tasks.

Furthermore, the MM-Verifier mechanism contributes significantly to the accuracy gains. Across all models, the transition from Majority Voting to MM-Verifier (Stage 2) consistently improves performance, underscoring the importance of verification-enhanced reasoning. In particular, MM-Reasoner benefits the most from this verification process, indicating that it effectively integrates verification feedback into its reasoning pipeline. Overall, these results validate the effectiveness of MM-Verify as a strong algorithm.


Input <question, solution>		
<p>Q: what is the total volume of the measuring cup?</p> <p>A: Let's think step by step.</p> <p>The answer is 1000.</p>		
Verify output from MM-Verifier		
<p>To verify the solution, let's analyze the given image and the solution step by step.</p> <p>1. Identify the Measuring Cup:</p> <ul style="list-style-type: none">- The image shows a transparent measuring cup with markings for volume in milliliters (mL) and grams (g). <p>2. Observe the Markings:</p> <ul style="list-style-type: none">- The visible markings on the cup include the numbers 0, 200, 400, 600, 800, and 1000. <p>3. Check the Solution:</p> <ul style="list-style-type: none">- The solution states that the total volume is 1000 g, which corresponds to 1000 mL based on the markings on the cup. <p>4. Expected Answer:</p> <ul style="list-style-type: none">- The expected answer is 1000 mL, which matches the solution. <p>5. Conclusion:</p> <ul style="list-style-type: none">- The solution correctly identifies the total volume of the measuring cup as 1000 mL. <p>Verification: Is the answer correct (Yes/No)? Yes</p>		
Verify output from Qwen2-VL-72B-Instruct		
No		

Figure 6: We present case of MM-Verifier. We can see MM-Verifier correctly verify the answer with CoT while Qwen2-VL-72B-Instruct failed to.

C Prompts

This paper primarily focuses on three key prompts: (1) the prompt for generating Verify data (Figure 7), (2) the prompt (Figure 8) for distilling the QwQ-32B-Preview model after converting multimodal data into a textual format, and (3) the prompt for extracting answer data (Figure 9).

D More Case Study

As shown in Figure 6, our MM-Verifier is capable of verifying the correctness of a solution that provides only a simple answer by leveraging CoT approach. In contrast, Qwen2-VL-72B-Instruct incorrectly classifies the solution as correct, highlighting its limitations in reasoning-based verification.

Table 5: We compare sub-task’s performance of Qwen2-VL-Instruct-7B, LLaMA-3.2-11B-Vision with our MM-Reasoner on the MathVista benchmark. GPS: geometry problem solving; MWP: math word problem; FQA: figure question answering; TQA: textbook question answering; VQA: visual question answering.

Model	Method	GPS	MWP	FQA	TQA	VQA	ALL
<i>Sample 4</i>							
Qwen2-VL	Majority Voting	44.7	62.9	63.6	58.2	54.7	57.1
	Qwen2-VL-7B as Judgement	51.4	63.4	62.8	53.8	55.9	57.9
	MM-Verifier(Stage1)	51.0	65.1	62.5	63.3	52.0	58.8
	MM-Verifier(Stage2)	51.9	66.7	64.7	60.8	53.6	59.8
LLaMA-3.2-11B-Vision	Majority Voting	38.0	48.9	45.7	55.7	39.7	45.2
	Qwen2-VL-7B as Judgement	39.9	51.6	48.7	52.5	38.0	41.7
	MM-Verifier(Stage1)	42.3	54.8	51.7	57.0	45.3	50.0
	MM-Verifier(Stage2)	46.2	53.8	53.2	56.3	42.5	50.4
MM-Reasoner	Majority Voting	57.2	62.4	65.4	57.6	51.4	59.4
	Qwen2-VL-7B as Judgement	45.7	59.1	56.1	55.7	48.6	53.1
	MM-Verifier(Stage1)	51.0	65.1	65.1	61.4	54.2	59.6
	MM-Verifier(Stage2)	58.7	66.1	66.2	63.9	50.8	61.5
<i>Sample 8</i>							
Qwen2-VL	Majority Voting	55.8	65.1	65.8	62.0	55.3	61.1
	Qwen2-VL-7B as Judgement	49.0	60.2	59.1	52.5	49.7	54.5
	MM-Verifier(Stage1)	53.8	69.9	66.2	63.9	53.1	61.6
	MM-Verifier(Stage2)	52.4	70.4	69.5	62.7	55.3	62.5
LLaMA-3.2-11B-Vision	Majority Voting	43.8	52.7	50.6	57.0	38.0	48.3
	Qwen2-VL-7B as Judgement	39.9	51.6	48.7	52.5	38.0	46.1
	MM-Verifier(Stage1)	45.2	54.3	54.6	62.0	41.3	51.4
	MM-Verifier(Stage2)	44.7	55.4	55.0	63.9	42.5	52.1
MM-Reasoner	Majority Voting	57.2	66.7	66.2	64.6	55.3	62.2
	Qwen2-VL-7B as Judgement	47.6	52.2	59.1	56.3	51.4	53.6
	MM-Verifier(Stage1)	56.7	72.0	66.2	68.4	55.7	63.8
	MM-Verifier(Stage2)	60.0	73.1	68.8	67.7	55.9	65.3
<i>Sample 12</i>							
Qwen2-VL	Majority Voting	62.5	67.7	65.4	62.7	54.7	62.9
	Qwen2-VL-7B as Judgement	49.5	60.8	59.5	51.3	48.6	54.4
	MM-Verifier(Stage1)	56.7	69.9	69.9	65.2	54.7	63.7
	MM-Verifier(Stage2)	58.7	67.7	69.5	65.8	57.0	64.1
LLaMA-3.2-11B-Vision	Majority Voting	47.1	54.3	54.3	61.4	39.7	51.3
	Qwen2-VL-7B as Judgement	42.3	45.2	50.6	54.4	34.1	45.5
	MM-Verifier(Stage1)	50.5	57.5	58.0	65.8	43.6	55.0
	MM-Verifier(Stage2)	55.8	56.5	59.5	62.7	44.1	55.9
MM-Reasoner	Majority Voting	65.2	68.0	68.2	64.4	55.3	64.6
	Qwen2-VL-7B as Judgement	46.6	61.3	59.5	56.3	52.5	55.4
	MM-Verifier(Stage1)	61.1	72.0	70.3	64.6	53.1	64.8
	MM-Verifier(Stage2)	62.0	71.5	69.9	67.1	53.6	65.3

Table 6: Comparison of the performance of MCTS with Majority Voting and our MM-Verify on the MathVista benchmark.

Base Model	# size	Method	GPS	MWP	FQA	TQA	VQA	ALL
llama-3.2-11B-Vision	11B	MCTS	42.8	16.1	33.8	48.1	40.2	35.8
		Majority Voting	38.0	48.9	45.7	55.7	39.7	45.2
		MM-Verify	46.2	53.8	53.2	56.3	42.5	50.4
Qwen2-VL	7B	MCTS	53.4	47.3	51.7	55.7	43.6	50.4
		Majority Voting	44.7	62.9	63.6	58.2	54.7	57.1
		MM-Verify	51.9	66.7	64.7	60.8	53.6	59.8
LLaVA-OneVision	7B	MCTS	72.1	56.5	44.6	58.9	45.3	54.9
		Majority Voting	66.8	54.8	39.4	57.6	40.8	51.1
		MM-Verify	66.8	63.4	47.6	58.2	43.0	55.4

Verify Label Prompt
<p>Solve the math problems and provide step-by-step solutions, ending with "The answer is [Insert Final Answer Here]".</p> <p>When asked "Verification: Is the answer correct (Yes/No)?", respond with " Yes" or " No" based on the answer's correctness.</p> <p>When asked "Verification: Let's verify step by step.", verify every step of the solution and conclude with "Verification: Is the answer correct (Yes/No)?" followed by " Yes" or " No".</p> <p>Q: {data['question']}</p> <p>A: Let's think step by step.</p> <p>{data['solution']}</p>

Figure 7: Verify Label Prompt. This is the prompt we built with reference to Zhang et al. (2024c).

QwQ-32B-Preview Distill Prompt
<p>Your role as an assistant involves thoroughly exploring questions through a systematic long thinking process before providing the final precise and accurate solutions. This requires engaging in a comprehensive cycle of analysis, summarizing, exploration, reassessment, reflection, backtracing, and iteration to develop well-considered thinking process.</p> <p>Please structure your response into two main sections: Thought and Solution.</p> <p>In the Thought section, detail your reasoning process using the specified format:</p> <pre> """ < begin_of_thought > {thought with steps separated with "\n\n"} < end_of_thought > """ </pre> <p>Each step should include detailed considerations such as analysing questions, summarizing relevant findings, brainstorming new ideas, verifying the accuracy of the current steps, refining any errors, and revisiting previous steps.</p> <p>In the Solution section, based on various attempts, explorations, and reflections from the Thought section, systematically present the final solution that you deem correct. The solution should remain a logical, accurate, concise expression style and detail necessary step needed to reach the conclusion, formatted as follows:</p> <pre> """ < begin_of_solution > {final formatted, precise, and clear solution} < end_of_solution > """ </pre> <p>Now, try to solve the following question through the above guidelines:</p>

Figure 8: QwQ-32B-Preview Distill Prompt. This is the prompt we built with reference to Min et al. (2024).

Answer Extract Prompt
<p>Hint: Please answer the question requiring an integer answer and provide the final value, e.g., 1, 2, 3, at the end.</p> <p>Question: Which number is missing?</p> <p>Model response: The answer is 14.</p> <p><Extracted answer>: 14</p> <p>...</p>
<p>Hint: Please answer the question requiring a Python list as an answer and provide the final list, e.g., [1, 2, 3], [1.2, 1.3, 1.4], at the end.</p> <p>Question: Between which two years does the line graph saw its maximum peak?</p> <p>Model response: The line graph saw its maximum peak between 2007 and 2008.</p> <p><Extracted answer>: [2007, 2008]</p>
<p>Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.</p> <p>Question: What fraction of the shape is blue?\nChoices:\n(A) 3/11\n(B) 8/11\n(C) 6/11\n(D) 3/5</p> <p>Model response: The correct answer is (B) 8/11.</p> <p><Extracted answer>: B</p>
<p>Please extract the following <Extracted answer> referencing above example. Do not output any other information.</p> <p>Question: {data['question']}</p> <p>Model response: {data['response']}</p> <p>Extracted answer:</p>

Figure 9: Answer Extract Prompt