UniMath-CoT: A Unified Framework for Multimodal Mathematical Reasoning with Re-Inference Affirmation

Zhixiang Lu[1], Mian Zhou[1], Angelos Stefanidis[1], Jionglong Su[1],

[1]School of Artificial Intelligence and Advanced Computing, Xi'an Jiaotong-Liverpool University

Correspondence: Jionglong.Su@xjtlu.edu.cn

Abstract

Large Language Models (LLMs) have achieved considerable success in text-based mathematical reasoning, yet their potential remains underexplored in the multimodal mathematics domain where joint text and image understanding is imperative. A key bottleneck hindering progress is the scarcity of high-quality, genuinely multimodal benchmarks. To address this gap, we construct a unified benchmark by consolidating and curating three public multimodal mathematics datasets. We subsequently propose the **UniMath-CoT** framework, which establishes a robust performance baseline by combining Chain-of-Thought (CoT) principles with efficient Supervised Fine-Tuning (SFT) based on Low-Rank Adaptation (LoRA). Furthermore, to bolster the model's reasoning robustness, we introduce an innovative verification mechanism, **AARI (Answer Affirmation by Re-Inference)**, which leverages a specialized re-inference protocol to have the model self-scrutinize and validate its initial conclusions. Our comprehensive experiments show that this integrated strategy substantially boosts performance, surpassing a wide range of open-source models and markedly closing the gap with leading proprietary systems.

1 Introduction

Mathematical reasoning is a cornerstone of human intelligence. Automating this process, particularly for problems presented in a multimodal format that integrates text with diagrams and figures, represents a significant frontier for artificial intelligence (Seo et al., 2015). This capability has profound implications for domains like personalized education, scientific discovery, and engineering, where complex information is often conveyed visually.

The advent of powerful Vision-Language Models (VLMs), such as GPT-4V, Gemini, and LLaVA (OpenAI, 2023; Team et al., 2023; Liu et al., 2024a), has opened new avenues for tackling this challenge. These models can, in principle, process interleaved text and images to produce human-like reasoning steps, moving beyond the text-only limitations of earlier models (Wei et al., 2022). However, the landscape of multimodal mathematical reasoning remains fragmented. Numerous datasets exist (Lu et al., 2024; Ling et al., 2023; Liu et al., 2024b; Wang et al., 2024), but they possess distinct formats, scopes, and a significant number of instances where the visual component is not essential for solving the problem. This fragmentation and lack of a unified, high-quality benchmark impede the rigorous evaluation of VLMs and hinder progress in the field. To address these challenges and systematically advance the state of multimodal mathematical reasoning, this paper makes three primary contributions:

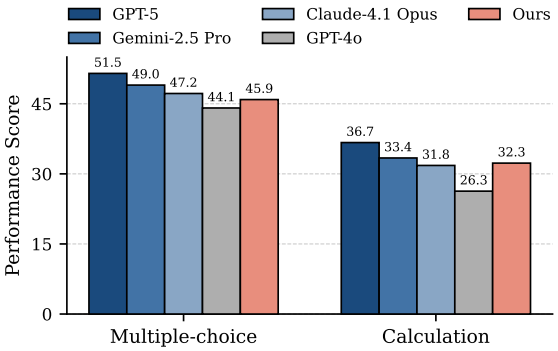

Figure 1: Performance comparison of our model against leading proprietary VLMs on Multiple-choice and Calculation problems. "Ours" refers to the Qwen2.5-VL 7B model fine-tuned with UniMath-CoT and enhanced with AARI.

1. **A Unified and Curated Multimodal Mathematics Dataset.** We construct a new benchmark by amalgamating three prominent datasets: MathViTa, MathVision, and CMM-MATH. We apply a rigorous filtering process to remove unimodal instances, ensuring that

Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP 2025), pages 176–185

every problem is genuinely multimodal. The resulting dataset serves as a more challenging and realistic testbed for evaluating VLM reasoning capabilities.

2. **An Effective Reasoning Framework (UniMath-CoT).** We leverage and validate UniMath-CoT, a structured Chain-of-Thought approach tailored for the complexities of joint visual and textual understanding. Implemented via parameter-efficient fine-tuning (LoRA) (Hu et al., 2021), this framework guides the model to generate more coherent and accurate reasoning paths than conventional methods.

3. **A Novel Answer Verification Technique (AARI).** We introduce AARI, a lightweight, post-hoc strategy to boost model performance. This technique directs the model to re-evaluate its predicted answer through a secondary, condensed inference pass, acting as a powerful self-correction mechanism that significantly reduces errors and improves final accuracy.

Table 1: Distribution of samples by split, source, and language. Each cell shows the absolute count and its proportion relative to the split's total count.

Split	Dataset	Language	
		Chinese	**English**
Train	MathVision	0 (0.0%)	3157 (23.8%)
	MathVista	357 (2.7%)	5428 (40.9%)
	EduChat-Math	4255 (32.1%)	61 (0.5%)
	Total	**4612 (34.8%)**	**8646 (65.2%)**
Test	MathVision	0 (0.0%)	187 (18.7%)
	MathVista	45 (4.5%)	311 (31.1%)
	EduChat-Math	455 (45.5%)	2 (0.2%)
	Total	**500 (50.0%)**	**500 (50.0%)**

2 Related Works

Multimodal Math Datasets The development of capable models for mathematical reasoning is intrinsically linked to the availability of high-quality datasets. Early efforts in this domain often focused on text-based problems, with benchmarks like GSM8K (Cobbe et al., 2021) and MathQA (Amini et al., 2019) becoming standard testbeds for evaluating the reasoning capabilities of LLMs.

The frontier has recently shifted towards multimodal problems that require understanding both text and images. This has led to the creation of several key benchmarks. For instance, Geometry3K (Lu et al., 2021) provides high-school level geometry problems with formally annotated diagrams. More recently, a new wave of diverse datasets has emerged, including **MathVista** (Lu et al., 2024), a comprehensive benchmark aggregating problems from 28 different sources; **MathViTa** (Ling et al., 2023), which focusing on visual instruction tuning for math; **MathVision** (Wang et al., 2024), designed to test reasoning-intensive math problems; and **CMM-MATH** (Liu et al., 2024b), a benchmark specifically for Chinese multimodal mathematics.

While these datasets have been invaluable, their varied formats, scopes, and annotation styles create a fragmented landscape. This makes it challenging to perform unified and fair evaluations of different models. Our work addresses this gap directly by curating and unifying three of these recent, diverse datasets into a single, filtered benchmark, ensuring all instances are genuinely multimodal and providing a more robust foundation for analysis.

Vision-Language Models (VLMs) The advent of powerful VLMs, pioneered by models like LLaVA (Liu et al., 2024a) and further advanced by open-source models like Qwen-VL (Bai et al., 2023) and proprietary systems like GPT-4V (OpenAI, 2023) and Gemini (Team et al., 2023), has enabled end-to-end multimodal reasoning. The primary challenge has since shifted to effectively eliciting their latent reasoning abilities.

The foundational technique for this is Chain-of-Thought (CoT)(Wei et al., 2022) prompting (Wei et al., 2022), which significantly improves performance by instructing models to generate step-by-step solutions. This paradigm has been extended to the multimodal domain, with methods like Multimodal-CoT (Zhang et al., 2023) that prompt the model to integrate information from both modalities in its reasoning steps. While effective, these zero-shot prompting methods can be sensitive to prompt formulation and may not be optimal for a specific domain. An alternative is to instill reasoning capabilities through training. Our *UniMath-CoT* framework aligns with this direction, employing supervised fine-tuning (SFT) (Ouyang et al., 2022) to teach the model a specialized reasoning structure for multimodal math, aiming for a more robust and replicable performance than prompting alone.

Self-Correction and Answer Verification A key limitation of LLMs, even when using CoT, is their propensity for making logical or computational errors within a reasoning chain. To mitigate this, research has explored methods for self-correction and answer verification (Lu et al., 2025; Li et al., 2025). One approach involves training a separate verifier model to score or judge the correctness of solutions generated by a primary model (Cobbe et al., 2021). Another popular direction is self-refinement, where the model iteratively critiques and improves its own output based on feedback (Madaan et al., 2023).

While powerful, these methods can be computationally expensive, requiring either the training of an additional model or multiple inference passes. Our proposed AARI technique is situated within this line of research but designed for efficiency. It is a lightweight, single-pass verification step that prompts the model to perform a final, focused check on its answer. AARI functions as a post-hoc self-correction mechanism that improves reliability without the overhead of multi-step refinement or external verifiers.

3 Methodology

Our methodology is designed to systematically enhance and evaluate the mathematical reasoning capabilities of Vision-Language Models (VLMs). It is founded on three core components: a newly curated benchmark, a fine-tuned reasoning framework, and a novel inference strategy for answer verification as shown in Figure 2.

3.1 UniMath Benchmark Construction

A robust evaluation requires a high-quality benchmark. To this end, we construct **UniMath**, a unified and curated benchmark derived from three recent and diverse multimodal math datasets: MathVista (Lu et al., 2024), MathVision (Wang et al., 2024), and CMM-MATH (Liu et al., 2024b). The construction process involves four critical steps:

1. **Schema Unification:** All problems from the source datasets are converted into a standardized JSON format. Each entry contains a unique ID, the raw question text, an image, the ground-truth answer, and, where available, human-annotated reasoning steps. This creates a consistent data structure for all subsequent processing.

2. **Answer Normalization:** We develop and apply a rigorous parsing function, $\phi(\cdot)$, to normalize all answers. This function extracts numerical values, correctly interprets fractions and percentages, removes units, and standardizes multiple-choice options (e.g., converting ‘(A)‘ to ‘A‘). This ensures that evaluation is based on mathematical correctness, not superficial formatting differences.

3. **Genuinely Multimodal Filtering:** A key contribution of UniMath is its focus on problems requiring genuine multimodal reasoning. We employ a systematic process to filter out instances where the image is redundant or extraneous, ensuring that a correct solution can only be derived by integrating both visual and textual information.

4. **Problem Type and Scope Curation:** To maintain evaluation clarity and objectivity, we further refine the benchmark by problem type and answer scope. Specifically, we only select two types of problems: multiple-choice questions and non-multiple-choice problems that possess a unique, definitive solution. Furthermore, to avoid ambiguity with extremely large numbers or complex symbolic expressions, we constrain the answers for all non-multiple-choice problems to be rational numbers within the range of `[-10,000, 10,000]`.

The final UniMath benchmark, shaped by this rigorous curation process, serves as the foundation for our experiments, providing a challenging and standardized testbed for multimodal mathematical reasoning.

3.2 UniMath-CoT Reasoning Framework

To move beyond the limitations of zero-shot prompting, we propose UniMath-CoT, a framework for teaching a VLM to generate structured, step-by-step reasoning for multimodal math problems through supervised fine-tuning (SFT).

The goal of UniMath-CoT is to train the model to produce a specific, decomposable reasoning chain that mirrors an ideal problem-solving process (see Figure 3). This chain consists of several stages: (1) *Visual Grounding*, where key information from the image is extracted; (2) *Problem Formulation*, where visual and textual information are integrated; (3) *Step-by-Step Planning*; (4) *Execution* of the plan with calculations; and (5) the *Final Answer*.

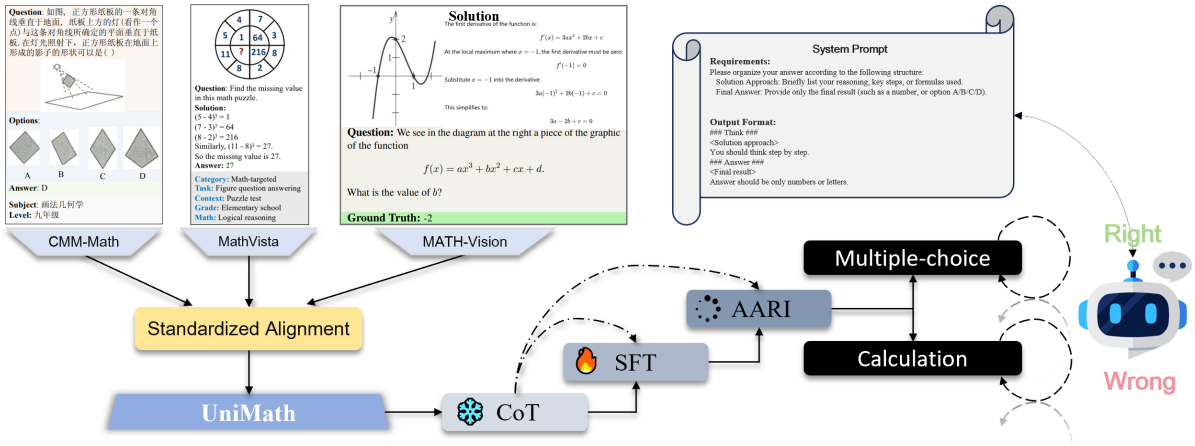

Figure 2: The overall framework of our work, encompassing the entire pipeline from data construction to model training and inference-verification. We first integrate three public datasets (CMM-Math, MathVista, MATH-Vision) through standardized alignment to build the unified **UniMath** benchmark. Subsequently, this benchmark is used to generate CoT formatted samples for SFT of the base VLM. During inference, the preliminary answer from the SFT model is passed through our proposed **AARI** module for re-inference and verification to produce a more reliable final result.

UniMath-CoT for Multimodal Mathematical Reasoning

Input: You are an expert skilled in mathematical reasoning. Your task is to analyze the given math problem and its type, and then provide the solution. The problem type is: {Question type} {Question image}
Requirements:
1. All mathematical symbols and formulas must be written using LaTeX syntax.
2. Absolutely no factual mistakes are allowed (for example, errors like $\ln 1 = 1$ will result in termination of cooperation).
3. Please organize your answer according to the following structure:
Solution Approach: Briefly list your reasoning, key steps, or formulas used.
Final Answer: Provide only the final result (such as a number, or option A/B/C/D).
Output Format:
Think
<Solution approach (in LaTeX format)>
- Step 1: $formula/derivation$
- Step 2: $formula/derivation$
- ...
■ Note: If it is a multiple-choice question, include the verification process. For example, if you derive that the answer is 20 and option B is 20, then the final Answer must be "B".
Answer
<Final result, only numbers or letters>
IMPORTANT:
- If the final answer is a number, it must be output as a floating-point number, not as an expression or fraction (e.g., `$\frac$`, `{frac} = 0.5`, or `1/2` are forbidden).
- If the answer is for a multiple-choice question, it must be a capital letter such as "C", not "C.9".
- Violating these rules in the final answer will result in termination of cooperation!
[Examples]
■ Calculation question (correct output):
Think
...(derivation leading to 1)
Answer
1
■ Multiple-choice question (correct output):
Think
...(derivation showing B is correct)
Answer
B
Now strictly follow these rules to handle the problem.

Figure 3: The structured reasoning generation prompt template for UniMath-CoT.

Formally, we define a problem instance as a tuple $P = (I, Q)$, where I represents the image and Q is the corresponding textual question. The desired output is a coherent reasoning chain $R = (s_1, s_2, \ldots, s_n)$ that culminates in the final answer A. For training purposes, we concatenate these elements into a single target sequence $Y = (s_1, \ldots, s_n, A)$, which the model must learn to generate autoregressively.

Our primary objective is to fine-tune a base Vision-Language Model (VLM), parameterized by θ, to maximize the conditional likelihood of generating this ground-truth sequence Y given the problem instance P. To achieve this in a computationally efficient manner, we employ Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning.

The training process minimizes the standard cross-entropy loss, which is equivalent to minimizing the negative log-likelihood of the target sequence. The loss function, $\mathcal{L}_{\text{SFT}}$, over our curated training dataset $\mathcal{D}_{\text{train}}$ is formally expressed as the expectation of this loss across all data samples:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(P,Y)\sim\mathcal{D}_{\text{train}}} \left[\log p_\theta(Y|P)\right] \quad (1)$$

where the log-likelihood of a single sequence is decomposed autoregressively as:

$$\log p_\theta(Y|P) = \sum_{t=1}^{|Y|} \log p_\theta(y_t|y_{<t}, P) \quad (2)$$

Algorithm 1 Answer Affirmation by Re-Inference

Require: Image I, question Q, initial answer A, prompt P, VLM $\mathcal{M}$, mode $\mu \in \{t_1, t_2\}$
Ensure: Final answer A^*

if $\mu = t_1$ **then** ▷ Multiple-Choice
  $\mathcal{O} = \{O_1, \ldots, O_k\}$
  for $i = 1, \ldots, k$ **do**
    $P \leftarrow$ "Is 'O_i' correct for Q given I?"
    if $\mathcal{M}(P) =$ "Correct" **then**
      return O_i
  return Fallback($\mathcal{O}$)
else if $\mu = t_2$ **then** ▷ Calculation
  $\mathcal{A}_+ \leftarrow \emptyset$
  for $j = 1, \ldots, N$ **do**
    $P \leftarrow$ "Is 'A_j' correct for Q given I?"
    if $\mathcal{M}(P) =$ "Correct" **then**
      $\mathcal{A}_+ \leftarrow \mathcal{A}_+ \cup \{A_j\}$
  if $|\mathcal{A}_+| > 0$ **then**
    return $\arg\max_{A \in \mathcal{A}_+} \#\{\, i : A_i = A \,\}$
  else
    return Fallback($\mathcal{A}$)

This training paradigm compels the model not only to predict the final answer but also to articulate the underlying step-by-step derivation. By optimizing the entire reasoning path, this process endows the model with a specialized and robust capability for structured, multimodal problem-solving.

3.3 Answer Affirmation by Re-Inference

Even with a strong reasoning framework, VLMs can produce fallacious conclusions from otherwise sound reasoning chains. To address this, we introduce a novel lightweight inference technique: AARI, which is a two-stage process designed to verify and self-correct the model's initial conclusion.

Stage 1: Candidate Generation Given a problem $P = (I, Q)$, the fine-tuned model first generates a candidate solution, which includes the reasoning chain R_1 and a preliminary answer A_1. This is the standard generative process:

$$(R_1, A_1) = \underset{R,A}{\text{argmax}}\ P_\theta(R, A | I, Q) \tag{3}$$

Stage 2: Affirmation via Re-Inference Next, instead of immediately accepting A_1, we formulate a new verification prompt, P', which contains the original problem, the generated reasoning, and the candidate answer: $P' = (I, Q, R_1, A_1)$. We then task the model with assessing the validity of A_1 given R_1. Let V be a latent variable representing the affirmation of the answer, where $V = 1$ signifies correctness and $V = 0$ signifies an error. The model implicitly computes the posterior probability $P_\theta(V = 1|P')$. The final answer, A_f, is determined by this affirmation step:

$$A_f = \begin{cases} A_1 & \text{if } p_\theta(V = 1|P') > 0.5 \\ A_2 & \text{otherwise} \end{cases} \tag{4}$$

where $A_2 = \arg\max_A p_\theta(A|P', V = 0)$. In practice, this is implemented by prompting the model with a question like, "Based on the provided reasoning, is the answer 'A_1' correct? Re-examine the steps and confirm." If the model's response is affirmative, we accept A_1. If it is negative, we prompt it to provide the corrected answer, A_2. This re-inference step forces the model to perform a final, focused consistency check, effectively reducing unforced errors without requiring external verifiers or complex iterative refinement.

4 Experiments

4.1 Experimental Setup

We conduct a comprehensive set of experiments to evaluate our proposed framework. All evaluations are performed on our newly constructed UniMath benchmark, which covers a diverse range of multimodal mathematics problems. The primary goals are to (1) assess the individual contributions of the UniMath-CoT fine-tuning strategy and the AARI inference mechanism through ablation studies, and (2) compare our final model's performance against state-of-the-art open-source and proprietary baselines. Our implementation is built upon PyTorch 2.11+. We leverage the Transformers library (v4.53+) for handling the underlying Vision-Language Models. To manage the significant computational requirements of fine-tuning, we employ `DeepSpeed` (Rajbhandari et al., 2020) for distributed training and memory optimization. For our parameter-efficient fine-tuning approach, we use `LoRA` with 64 lora rank, which is implemented with the `bitsandbytes` library (v0.42). The development environment requires `Python 3.11` and `CUDA 12.6`. All fine-tuning experiments were conducted on a server equipped with NVIDIA A100 80GB GPUs, a 32-core AMD EPYC processor, and 128GB of DDR4 memory. This configuration supported a per-device batch size of 8 during the LoRA fine-tuning process.

Table 2: Comprehensive benchmark on the UniMath dataset. We first establish a zero-shot performance leaderboard across proprietary and open-source models (<10B). We then conduct a detailed ablation study on the top-performing open-source model, Qwen2.5-VL, demonstrating the progressive performance gains from Chain-of-Thought (CoT), a standard SFT (LoRA), and finally our AARI method. Our approach, highlighted in gray and marked with a star (★), achieves state-of-the-art results among its peers, rivaling top proprietary systems.

Model / Method	Problem Type		Language		Overall
	Multiple-choice	Calculation	Chinese	English	Accuracy (%)
Proprietary Models (Zero-shot)					
GPT-5 (OpenAI, 2024)	**51.5**	**36.7**	**43.5**	**44.7**	**44.1**
Gemini-2.5 Pro (Reid and Team, 2024)	49.0	33.4	40.1	42.3	41.2
Claude-4.1 Opus (Anthropic, 2024)	47.2	31.8	38.8	40.2	39.5
GPT-4o (OpenAI, 2023)	44.1	26.3	33.0	37.4	35.2
Open-Source Models (Zero-shot)					
InternVL3 8B (Chen et al., 2024)	29.1	13.1	21.3	20.9	21.1
+ CoT Prompting	35.0	19.8	28.1	26.7	27.4
+ SFT (LoRA)	36.5	21.9	29.5	28.9	29.2
Qwen2.5-VL 7B (Bai et al., 2024)	30.5	13.7	23.8	20.4	22.1
+ CoT Prompting	35.3	19.5	28.2	26.6	27.4
+ SFT (LoRA)	37.2	21.5	31.4	30.9	31.2
DeepSeek-VL 7B (DeepSeek, 2024)	28.8	11.7	20.9	19.7	20.3
+ CoT Prompting	34.8	19.2	28.0	25.9	27.0
+ SFT (LoRA)	35.8	20.4	28.3	27.9	28.1
Our Method (Based on Qwen2.5-VL 7B)					
+ UniMath-CoT with SFT (LoRA)	39.6	26.8	34.5	31.9	33.2
★ + AARI	45.9	32.3	40.2	38.0	39.1

4.2 Evaluation Metric

Our primary evaluation metric is `Accuracy`, defined as the percentage of correctly solved problems. An answer is considered correct if the model's prediction, after applying a normalization function $\phi(\cdot)$, exactly matches the normalized ground-truth answer. Formally, for a test set $\mathcal{D}_{\text{test}}$, accuracy is defined as:

$$\text{Accuracy} = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{i=1}^{|\mathcal{D}_{\text{test}}|} \mathbb{I}(\phi(\hat{A}_i) = \phi(A_i)) \quad (5)$$

where $\hat{A}_i$ is the predicted answer, A_i is the ground-truth answer, and $\mathbb{I}(\cdot)$ is the indicator function.

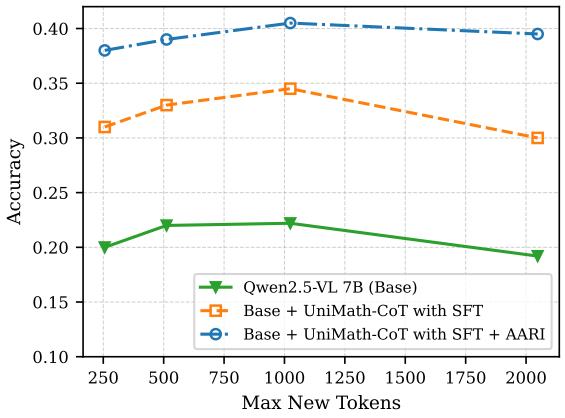

Figure 4: Performance of different model versions under varying maximum new token lengths.

4.3 Results and Analysis

Overall Performance As shown in Table 2, our full approach that combining the UniMath-CoT fine-tuning with the AARI inference strategy, achieves a final accuracy of 39.1%. This result establishes a new state-of-the-art among open-source models on this challenging task. Notably, our model significantly surpasses the powerful proprietary model GPT-4o (35.2%) and becomes highly competitive with Claude-4.1 Opus (39.5%).

The Impact of Each Component To isolate the contributions of our proposed methods, we conducted a detailed ablation study. We evaluated the effectiveness of our fine-tuning framework. Starting from a standard SFT Qwen2.5-VL 7B model which achieves 31.2% accuracy, employing our UniMath-CoT framework boosts the performance to 33.2%, which validates that our specialized approach of training on structured, decomposable

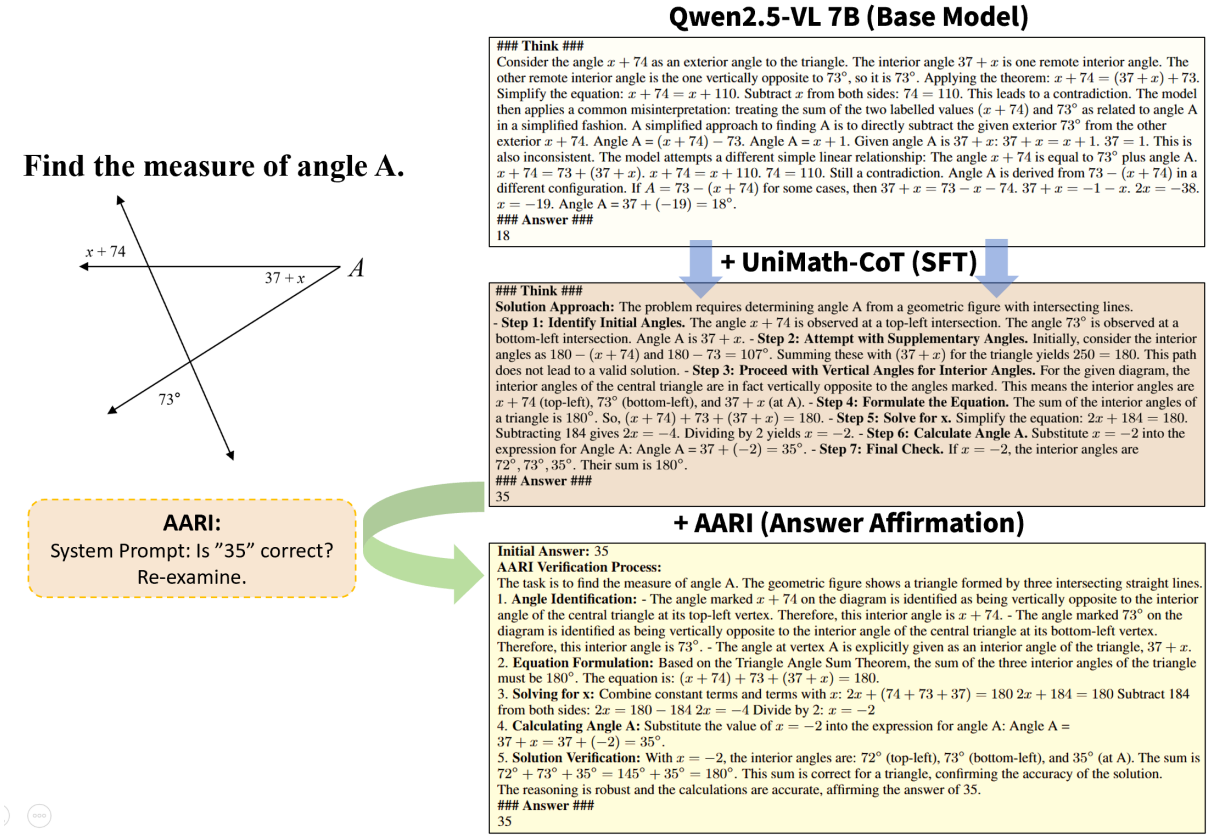

Figure 5: Qualitative Comparison of Model Outputs on a Sample Problem.

reasoning paths is more effective than generic instruction tuning. We also assessed the impact of our inference strategy. Applying AARI on top of the UniMath-CoT model elevates the accuracy from 33.2% to 39.1%. This represents a remarkable +5.9% increase, which corresponds to a 17.8% relative error reduction, underscoring the efficacy of the self-verification mechanism. The benefits of AARI are robust and consistent across all problem categories, boosting accuracy on multiple-choice (+6.3%) and calculation (+5.5%) problems, as detailed in Table 2. This demonstrates that AARI is a broadly applicable technique that enhances model reliability.

Effect of Generation Length We also investigated the impact of the maximum generation length (`Max New Tokens`) on performance. Figure 4 reveals two key findings. First, the performance hierarchy remains consistent across all token lengths, with the base model being outperformed by UniMath-CoT, which is in turn outperformed by the full model with AARI. This visually confirms our ablation results. Second, all model variants achieve their peak performance around 1024 max new tokens, suggesting this length provides an optimal balance between allowing for complete reasoning and avoiding excessive, potentially noisy, output.

Comprehensive Performance Profile To provide a more holistic, multi-dimensional view of our model's capabilities, we present a comparative radar chart in Figure 6. This chart visualizes the trade-offs between model size, inference speed, and performance on different sub-tasks (Chinese, English, Multiple-choice, Calculation) for our model and the baselines.

The chart clearly illustrates the well-rounded and highly efficient profile of our approach. Compared to its base model, **Qwen2.5-VL 7B**, our model shows a significantly larger and more balanced performance polygon. This expansion across all accuracy axes: Chinese, English, Multiple-choice, and Calculation, visually confirms that the gains from UniMath-CoT and AARI are comprehensive and not limited to a single domain.

When compared against leading proprietary models, our model demonstrates remarkable competitiveness despite its significantly smaller parameter size. For instance, while models like **GPT-5** and **Gemini-2.5 Pro** exhibit the largest perfor-

mance polygons overall, our 7B model achieves an accuracy profile that is notably competitive with, and in some areas superior to, much larger models like **GPT-4o**. This highlights the efficiency of our approach: we have successfully closed a substantial portion of the performance gap with state-of-the-art systems while operating at a fraction of the computational scale. The radar chart thus underscores our primary contribution: a clear and effective methodology for developing highly capable and efficient open-source models for complex multimodal reasoning.

Qualitative Analysis To showcase the differences in multi-step reasoning, Figure 5 provides a qualitative comparison of different models on a geometry problem that requires robust spatial interpretation. As shown, the base model fails due to a misinterpretation of geometric relations, while the UniMath-CoT model, despite an initial flawed step, successfully self-corrects to find the correct solution (Wei et al., 2022). This comparison highlights the critical role of structured reasoning and verification mechanisms, like those in the CoT and AARI models, in achieving reliable and accurate mathematical problem-solving.

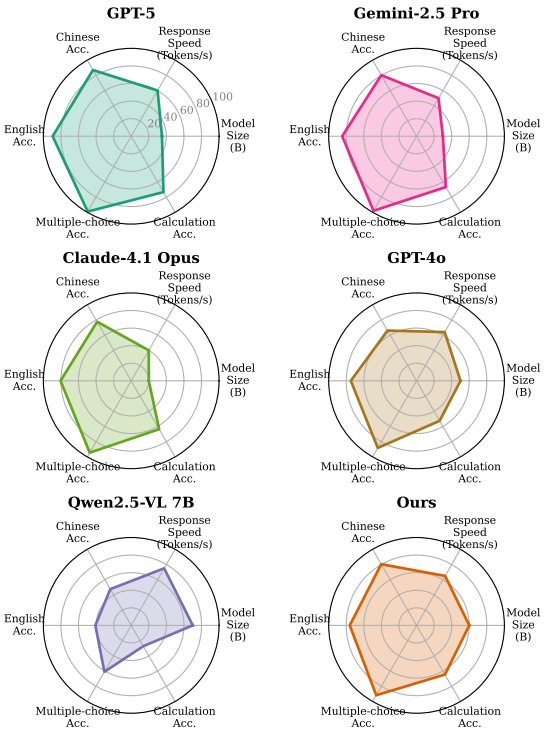

Figure 6: Multi-dimensional performance radar chart comparison. (All accuracy scores are multiplied by 2; parameter sizes for proprietary models are estimates.)

5 Conclusion and Limitation

In this work, we addressed key obstacles in multimodal mathematical reasoning, namely benchmark fragmentation and the need for robust verification. We introduced a tripartite contribution: (1) UniMath, a unified and rigorously curated benchmark for genuinely multimodal math problems; (2) UniMath-CoT, a fine-tuning framework that teaches structured reasoning; and (3) AARI, a novel, lightweight inference-time strategy for answer verification. Our experiments validate the power of this integrated approach. Our 7B model, enhanced by UniMath-CoT and AARI, achieves 39.1% accuracy on the UniMath benchmark, setting a new state-of-the-art for open-source models and outperforming strong proprietary systems like GPT-4o. A key finding is the remarkable effectiveness of AARI, which alone contributes a 5.9 % improvement, drastically reducing errors through its efficient self-verification mechanism.

Our work provides a clear roadmap for building powerful, open-source reasoning systems that rival proprietary models. Crucially, AARI demonstrates that inference-time self-correction is a highly effective strategy for boosting model factuality and reliability, a principle with strong potential for generalization beyond mathematics to other complex domains. Future work can extend this foundation by exploring iterative self-correction mechanisms and expanding the UniMath benchmark to new modalities like video-based challenges.

Ethics Statement

Our work leverages Large Language Models (LLMs) for complex mathematical problem-solving, specifically geometric reasoning, rather than direct text generation. While this application domain typically presents fewer immediate ethical concerns related to content generation biases, it is crucial to acknowledge the broader ethical landscape of LLM usage. Recent investigations have indicated that advanced prompting techniques, such as Chain-of-Thought (CoT) prompting, may inadvertently introduce or amplify ethical biases within LLM reasoning processes, even in non-generative tasks (Shaikh et al., 2023). Therefore, future work will involve a thorough examination of potential biases that might emerge from our method's reliance on CoT and answer affirmation techniques, ensuring fairness and robustness in mathematical reasoning applications.

References

Aida Amini, Mark Hopkins, Tony Tung, Cissy Le, Michael Huth, and Richard Socher. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2957–2967.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. https://www.anthropic.com/news/the-claude-3-model-family-opus-sonnet-haiku.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Jinze Bai, Zhenting Tu, Shusheng Yang, Qian Ma, Yichang Zhu, Shijie Wang, Peng Wang, Sinan Tan, and Chang Zhou. 2024. Qwen2: A family of powerful open-source language models. *arXiv preprint arXiv:2406.04741*.

Zhe Chen, Weiyun Zhang, Wen Wang, Yiliang Liu, Zhaoyang Zhang, Jian Wang, Jie Luo, Yu Qiao, and Wenhai Wang. 2024. Internvl 1.5: A general vision-language model. *arXiv preprint arXiv:2404.16821*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

DeepSeek. 2024. Deepseek-vl: Towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Yulong Li, Zhixiang Lu, Feilong Tang, Simin Lai, Ming Hu, Yuxuan Zhang, Haochen Xue, Zhaodong Wu, Imran Razzak, Qingxia Li, and 1 others. 2025. Rhythm of opinion: A hawkes-graph framework for dynamic propagation analysis. *arXiv preprint arXiv:2504.15072*.

He Ling, Sen Liu, Jian Liu, Yixuan Li, Guangda Shi, Leyang Zhou, Zhaoyang Hu, Yixuan Sun, Xing Su, Jiaxuan Yu, and 1 others. 2023. Mathvita: A visual instruction tuning toolkit for general-purpose multimodal llms. *arXiv preprint arXiv:2308.03720*.

Haotian Liu, Chunyuan Li, Feiyu Li, and Michihiro Yasunaga. 2024a. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Wentao Liu, Qianjun Pan, Yi Zhang, Zhuo Liu, Ji Wu, Jie Zhou, Aimin Zhou, Qin Chen, Bo Jiang, and Liang He. 2024b. Cmm-math: A chinese multimodal math dataset to evaluate and enhance the mathematics reasoning of large multimodal models. *Preprint*, arXiv:2409.02834.

Pan Lu, Swaroop Mishra, Tony Xia, Liangcheng Huang, Ramana Al-Tawy, Yixin Jia, Wen-haw Zhang, Ping Yang, Mohamed Abdel-Hady, Bassem Majumder, and 1 others. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Pan Lu, Ran Zhang, Jivko Sashank, AI Wang, Mohit Singh, Hongxin Zhu, and Yi Su. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2028–2041.

Zhixiang Lu, Peichen Ji, Yulong Li, Ding Sun, Chenyu Xue, Haochen Xue, Mian Zhou, Angelos Stefanidis, Jionglong Su, and Zhengyong Jiang. 2025. Advancing low-resource machine translation: A unified data selection and scoring optimization framework. In *Advanced Intelligent Computing Technology and Applications*, pages 482–493, Singapore. Springer Nature Singapore.

Aman Madaan, Niket Tandon, Prakhar Gupta, Kevin Hall, Luyu Gao, Sreyan Majumder, Julian McAuley, Yiming Narayan, and Gena Sim. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

OpenAI. 2023. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf.

OpenAI. 2024. Gpt-4o. https://openai.com/index/hello-gpt-4o/.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. *Preprint*, arXiv:1910.02054.

M. Reid and Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Google AI*.

Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. 2015. Solving geometry problems: Combining text and diagram interpretation. *arXiv preprint arXiv:1509.04232*.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2023. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4454–4470, Toronto, Canada. Association for Computational Linguistics.

Gemini Team and 1 others. 2023. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ben Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Zhuosheng Zhang, Aston Li, Vladislav Lialin, Hai Lu, Hanjung Lee, Mohammad Hosseini, Xiang Li, Haotian Liu, Chunyuan Li, Mu Li, and Anima Anandkumar. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.