SpatialMath: Spatial Comprehension-Infused Symbolic Reasoning for Mathematical Problem-Solving

Ashutosh Bajpai[1,*], Akshat Bhandari[1,*], Akshay Nambi[3], Tanmoy Chakraborty[1,2]

[1]Indian Institute of Technology Delhi, India
[2]Indian Institute of Technology Abu Dhabi, UAE
[3]Microsoft Research, India
{eez228482,tanchak}@ee.iitd.ac.in, *akshatbhandari15@gmail.com*,
akshayn@microsoft.com

Abstract

Multimodal Small-to-Medium sized Language Models (MSLMs) have demonstrated strong capabilities in integrating visual and textual information but still face significant limitations in visual comprehension and mathematical reasoning, particularly in geometric problems with diverse levels of visual infusion. Current models struggle to accurately decompose intricate visual inputs and connect perception with structured reasoning, leading to suboptimal performance. To address these challenges, we propose `SpatialMath`, a novel `Spatial Comprehension-Infused Symbolic Reasoning Framework` designed to integrate spatial representations into structured symbolic reasoning chains. `SpatialMath` employs a specialized perception module to extract spatially-grounded representations from visual diagrams, capturing critical geometric structures and spatial relationships. These representations are then methodically infused into symbolic reasoning chains, facilitating visual comprehension-aware structured reasoning. To this end, we introduce `MATHVERSE-PLUS`, a novel dataset containing structured visual interpretations and step-by-step reasoning paths for vision-intensive mathematical problems. `SpatialMath` significantly outperforms strong multimodal baselines, achieving up to 10 percentage points improvement over supervised fine-tuning with data augmentation in vision-intensive settings. Robustness analysis reveals that enhanced spatial representations directly improve reasoning accuracy, reinforcing the need for structured perception-to-reasoning pipelines in MSLMs.

1 Introduction

Recent progress in generative AI has produced Multimodal Small-to-medium sized Language Models (MSLMs) capable of integrating visual and textual information for tasks such as image captioning and multimodal question answering. However, MSLMs perform poorly on mathematical problems requiring visual comprehension, including geometric diagrams and spatial reasoning, due to their vision encoders being optimized for general image understanding rather than for capturing the structured, symbolic, and spatial information essential for mathematical tasks. This limits their effectiveness in STEM applications and mathematical problem solving.

Do MSLMs Truly Comprehend Visual Mathematical Information? The performance gap between text-based and visually presented math problems raises a deeper question: *Do MSLMs actually comprehend mathematical information from visual inputs, or are they simply generating approximations based on weak feature extraction?* While recent MSLMs such as InternLM-XC2 (Dong et al., 2024), LLaVA-NeXT (Liu et al., 2024b), LLaVA-OneVision (Li et al., 2024), and Phi-3 (Abdin et al., 2024b) excel in general vision-language tasks, including image captioning (Vinyals et al., 2015; Ge et al., 2024) and document-based question answering, they struggle with math reasoning tasks that rely on purely visual representations. This limitation is not merely anecdotal – quantitative evaluations from MathVerse, a recently introduced benchmark for assessing visual mathematical reasoning in MSLMs, highlight these deficiencies (Zhang et al., 2024a). MathVerse evaluates models across varying levels of textual-visual integration (Figure 1), revealing that while MSLMs perform well on text-heavy problems, they fail on vision-dense tasks like diagram-based angle computation and coordinate-based proofs. For example, LLaVA-NeXT-34B achieves 20% on text-dominant problems but only 9% on vision-only tasks. Furthermore, another similar study, the DYNAMATH (Zou et al., 2025) benchmark, revealed that VLMs consistently produce identical answers despite vari-

[*]Equal Contribution.

Findings of the Association for Computational Linguistics: EACL 2026, pages 717–742

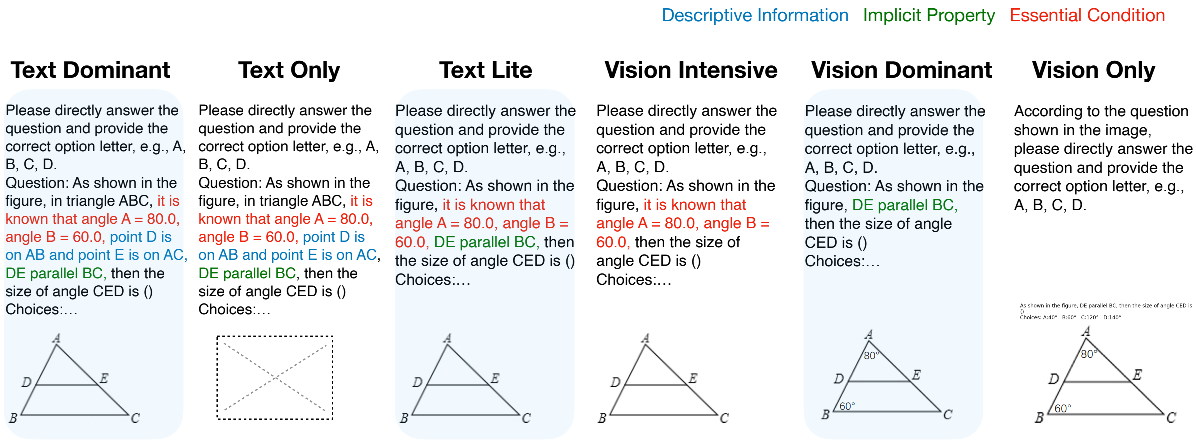


Figure 1: An example from our dataset showcasing six settings of a geometrical mathematical problem.

ations in visual conditions. This behavior suggests a reliance on memorization rather than reasoning grounded in generalized underlying principles. This highlights the need to bridge visual perception with mathematical understanding.

Given the observed deficiencies in MSLMs' ability to process and reason about visually complex mathematical problems, we investigate the following key research questions:

- **RQ1:** How effectively do small-to-medium sized multimodal language models (MSLMs) extract and understand structured information from visual mathematical problems?

- **RQ2:** Given an enriched textual spatial representation of a mathematical problem, can small-to-medium sized multimodal language models (MSLMs) generate a structured, stepwise reasoning path leading to the correct solution?

To overcome the aforementioned limitations of current MSLMs in mathematical reasoning, we introduce `SpatialMath` that explicitly infuses visual comprehensions into symbolic reasoning. Our approach consists of:

- `SpatialMath-SX`: Spatial Comprehension Core. A vision-language model optimized for extracting structured comprehension from visual mathematical content. We formulate `SpatialMath-SX` that utilizes the structured interpretation generated by an MLLM, also referred as spatial comprehension (SC), during the fine-tuning stage. MSLMs face challenges with generic vision encoders that lack specialization for mathematical content. To address this, we employ Spatial Comprehension Learning, where `SpatialMath-SX` transforms visual inputs into enhanced textual interpretations, capturing symbolic dependencies and spatial relationships.

- `SpatialMath-RX`: Reasoning Infusion Core. A dedicated solver that integrates spatial comprehension with symbolic reasoning to generate a logically coherent solution. `SpatialMath-RX` produces step-by-step solutions based on spatial comprehension produced by `SpatialMath-SX`. It emphasizes mapping reasoning chains (r) for logical consistency and symbolic coherence instead of relying solely on pattern matching. This approach enhances interpretability and accuracy by integrating visual perception with formal mathematical reasoning.

We conduct comprehensive experimentation with each module of `SpatialMath`. Notably, `SpatialMath` achieves a mean accuracy enhancement of 10.2% and 3.8% over the standard baseline of supervised fine-tuning with data augmentation on LLaVA-NeXT-34B and Phi-4, respectively. Furthermore, `SpatialMath-SX` achieves a mean accuracy enhancement of 2.0% (range: 0.8%-4.8%) across eight default MSLMs on varying degrees of visual infusion in a problem relative to the default performance. Our contributions are as follows[1]:

- Our research shows that MSLMs struggle significantly with understanding and reasoning through visually intricate mathematical problems, resulting in less effective problem-solving.

- To address these challenges, we propose `Spatial Comprehension-Infused Symbolic Reasoning Framework` (`SpatialMath`), a novel methodology to enhance both visual comprehension and symbolic reasoning, supported by

[1]Source code and dataset are available at https://github.com/ab-iitd/spatial-math

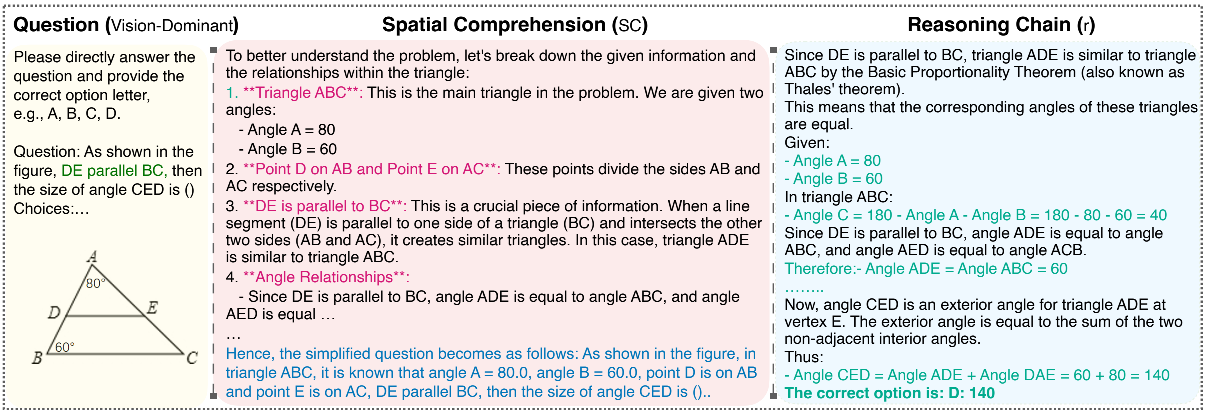


Figure 2: An instance of SC: the Spatial Comprehension generated by an MLLM, specifically GPT-4o that includes the text-only variant of the problem highlighted in with blue color, and r: the Reasoning Chain, which consists of a series of sequential steps culminating in the final solution, also generated by the aforesaid MLLM.

MATHVERSE-PLUS, a novel dataset that offers spatial comprehensions of complex visual math problems and structured reasoning paths across diverse visual complexity.

- Multi-faceted empirical evaluations demonstrate that SpatialMath consistently outperforms existing baselines, improving both comprehension and problem-solving accuracy in visual mathematical reasoning across datasets and MSLMs.

2 Dataset

We introduce MATHVERSE-PLUS, an enhancement of the MathVerse dataset, designed to facilitate our SpatialMath. MathVerse is the only publicly available dataset that systematically integrates varying levels of visual and textual information into mathematical problems, making it a suitable foundation for evaluating MSLMs' multimodal reasoning capabilities. Originally derived from multiple public sources (Chen et al., 2021; Lu et al., 2021), MathVerse reformulates mathematical problems into six different representations – ranging from text-only to vision-only – capturing different degrees of visual infusion (see Figure 1). Building upon this foundation, MATHVERSE-PLUS extends MathVerse by enriching problem variants with additional structured descriptions and stepwise reasoning annotations as depicted in Figure 2. This augmentation ensures that MSLMs can learn both visual comprehension and symbolic reasoning more effectively. Data construction steps are presented in Algorithm 1.

Statistics and Quality of MATHVERSE-PLUS. In the construction of MATHVERSE-PLUS, this study focuses exclusively on 2D and 3D geometric problems, deliberately excluding issues related to functions to ensure the dataset is homogenized for geometry-related purposes. Additionally, we consider only five visual modalities derived from the source dataset: Text-Dominant, Text-Lite, Vision-Intensive, Vision-Dominant, and Vision-Only. The Text-Only modality has been excluded, as it does not align with the multimodal focus of this research. MATHVERSE-PLUS consists of a total of 2.76K training instances, uniformly distributed across all five settings. The test set comprises 500 uniformly distributed instances, reserved for evaluating model performance and reporting results.

We conduct a manual evaluation of randomly-selected 30 instances. Using a Likert scale (Likert, 1932) (0–5, with 5 indicating maximum correctness and coherence), we obtain mean scores of 4.36 ± 0.86 and 4.46 ± 1.49 for MLLM generated Spatial Comprehensions (SC) and Reasoning Chain (RC), respectively, demonstrating high reliability of the generated annotations (see Appendix A for more details on MATHVERSE-PLUS).

3 Proposed Methodology

3.1 Preliminaries

We define a dataset D consisting of M geometry questions and corresponding answers, represented as pairs $(q^j_{(t_{(d,i,e)},v)}, a^j)$, where $j \in \{1, 2, \ldots, |M|\}$. Each question q^j contains textual (t) and visual (v) components. The textual component t is structured into:

- **Descriptive Information** (d): Explicitly provided details, i.e., geometric labels and numerical values.
- **Implicit Properties** (i): Inferred relationships,

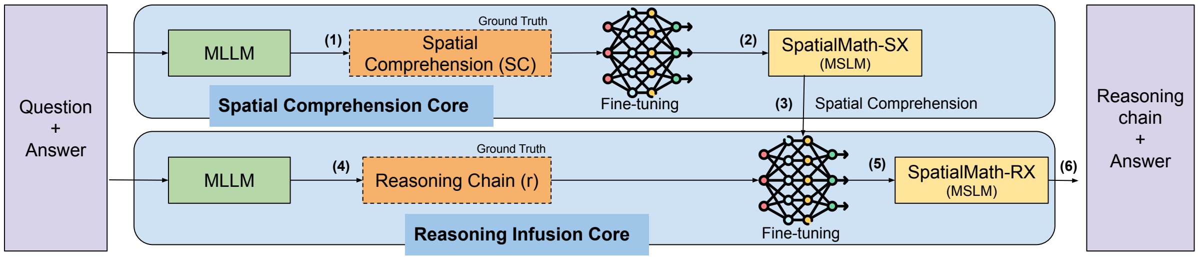


Figure 3: An overview of our framework: (1) Generating `Spatial Comprehension` (`SC`) for training set using an MLLM; (2) Fine-tuning `SpatialMath-SX` to enhance visual comprehension; (3) Using `SpatialMath-SX` at inference to generate `SC` for test set; (4) Generating `Reasoning Chain` data for train set; (5) Fine-tuning `SpatialMath-RX` using generated `SC` for training set by `SpatialMath-SX` and `Reasoning Chain`; (6) Deploying `SpatialMath-RX` at inference for problem-solving.

Algorithm 1 Data Construction Approach

Input: Dataset D containing M pairs of geometry questions and corresponding answers.

Step 1: Define Problem Variants.

- Define a set K consisting of six distinct problem variants, each with different levels of textual and visual integration.

Step 2: Generate Ground-Truth Spatial Comprehension.

- Introduce `Spatial Comprehension` (`SC`), a elaborative structured textual description of the problem's visual elements generated by an LLM serving as the ground-truth representation of the problem's visual content in text-only form.
- Enhance the dataset with richer linguistic interpretations.

Step 3: Enrich the Spatial Comprehension.

- Enrich the LLM's generated `Spatial Comprehension` (`SC`) with the text-only form of the problem.
- Ensure factual summary of text-based interpretation of visual elements.

Step 4: Generate Stepwise Reasoning Paths.

- For each problem variant, leverage a larger model to generate structured `Reasoning Chain` (`RC`), capturing stepwise logical dependencies and symbolic transformations.
- Use these structured reasoning paths to derive the correct solution and improve interpretability.

Output: A dataset enriched with textual representations, stepwise reasoning paths, and diverse problem variants.

including parallelism and symmetry.

- **Essential Conditions** (e): Key algebraic constraints necessary for solving the problem.

The objective is to predict a^j given the multimodal input $q^j_{(t_{(d,i,e)},v)}$, formulated as:

$$a^j = \arg\max_{a \in A} p\left(a \mid q^j_{(t_{(d,i,e)},v)}\right) \qquad (1)$$

where A is the answer space.

3.1.1 Problem Variants and Transformations

We extend the MathVerse formulation with varying textual and visual integration to five settings and exclude the Text-Only modality to align with the multimodal focus of this research.

- **Text-Dominant** ($q^j_{(t_{(d,i,e)},v)}$): All textual and visual input.
- **Text-Lite** ($q^j_{(t_{(i,e)},v)}$): Text lacks descriptive details (d), requiring reliance on i and e.
- **Vision-Intensive** ($q^j_{(t_{(e)},v)}$): Only essential conditions (e) are retained in text; implicit properties (i) must be inferred from the visual input.
- **Vision-Dominant** ($q^j_{(t_{(i)},v')}$): Only implicit properties (i) in Text, while numerical values are visually embedded.
- **Vision-Only** ($q^j_{(,v'')}$): The problem is presented purely visually, with no textual input.

Here, v' enriches the diagram with essential numeric labels, while v'' is fully self-contained, requiring complete visual parsing.

3.2 `SpatialMath-SX`

While MSLMs aim to jointly comprehend and solve visually complex mathematical problems, their visual comprehension capabilities remain a bottleneck. To address this, we introduce `SpatialMath-SX`: Spatial Comprehension Core, a dedicated module that independently evaluates and enhances the model's ability to interpret visual information before reasoning over it.

3.2.1 Spatial Comprehension

`SpatialMath-SX` is trained on an auxiliary dataset D', derived from D, where each question-answer pair (q^j, a^j) in D is transformed into a question-comprehension pair (q^k, c^{sc}) in D'. Here, c^{sc} represents spatial comprehension of the question, extracted from an MLLM that explicitly describes key visual and textual elements. The transformation allows `SpatialMath-SX` to focus solely on the comprehension, independent of reasoning.

The objective of `SpatialMath-SX` is to maximize the likelihood of generating the most accurate spatial comprehension `SC` given the multimodal question q^k, is presented in Equation 2, where B denotes the vocabulary space of possible comprehensions. Additionally, to enrich the extracted comprehension, we integrate text-only variant of the problem with LLM generated SC, resulting in the enhanced objective:

$$\mathtt{SC}^k_{t_{(d,i,e)}} = \arg\max_{sc \in B} p\left(sc \mid q^k_{(t^*,v^*)}\right) \quad (2)$$

3.2.2 Learning Objective

Many mid-sized multimodal models struggle with effective visual comprehension. To improve this, we fine-tune a dedicated SpatialMath-SX using parameter-efficient fine-tuning (PEFT) (Han et al., 2024; Hu et al., 2021). Specifically, we apply Low-Rank Adaptation (LoRA) to the multi-layer perceptron (MLP) projector within the pre-trained LLaVA-NeXT-34B model while keeping the LLM and vision encoder frozen. A similar setup is followed for Phi-4.

SpatialMath-SX is trained by minimizing the negative log-likelihood (NLL) loss across all training instances n':

$$\mathcal{L}_{SX} = -\sum_{k=1}^{n'} \log P(\mathtt{SC}^k_{t_{(d,i,e)}} \mid q^k_{(t^*,v^*)}) \quad (3)$$

3.3 SpatialMath-RX

An SpatialMath-SX can interpret the visual aspects of math problems and bridge different problem formats in vision-only settings but does not improve step-wise reasoning toward the correct solution. Mid-sized models often struggle with multi-step deduction, and multimodal objectives rarely align spatial features with structured reasoning. To address this, we propose SpatialMath-RX: Reasoning Infusion Core, which embeds spatial comprehension into reasoning chains for context-aware, step-wise symbolic reasoning.

3.3.1 Learning Objective

To this end, we leverage the Reasoning Chain (RC) generated by a comparatively larger model to enhance the performance of a medium-sized solver. The solver model is designed to first generate a Reasoning Chain, r^k, then derive the final answer a^k given $q^k_{(t^*,v^*)}$ and SC, $(\mathtt{SC}^k_{t_{(d,i,e)}}$. Unlike SpatialMath-SX, which is trained using structured spatial comprehension labels (SC), SpatialMath-RX is fine-tuned using a dataset explicitly enriched with structured reasoning steps (RC) to facilitate progressive mathematical inference.

$$\mathcal{L}_{RX} = -\sum_{k=1}^{n'} \log P(r^k, a^k \mid q^k_{(t^*,v^*)}, \mathtt{SC}^k_{t_{(d,i,e)}}) \quad (4)$$

In the process, an explicit alignment is achieved between spatial comprehension and the structured symbolic reasoning steps. The enhanced comprehension is leveraged in the reasoning infusion core (SpatialMath-RX), which builds upon SpatialMath-SX's extracted information to perform accurate problem-solving. We employ the LLaVA-NeXT-34B and Phi-4 models as the base models, utilizing a similar PEFT strategy as SpatialMath-SX. The joint optimization for reasoning follows in Equation 4.

Finally, the summarized formulation SpatialMath is composed of both the spatial comprehension core and the reasoning infusion core (SpatialMath-SX + SpatialMath-RX) as shown in Figure 3:

4 Experimental Setup

We conduct rigorous experiments using LLaVA- and Phi-based architectures to evaluate the proposed method. The dataset is split into 80% train and 20% test sets. We use GPT-4o to generate comprehension cues and reasoning chains for data augmentation. All hyperparameters, prompts, and other details are provided in the Appendix D.1 and Appendix D.2, respectively.

Evaluation Metrics. We use accuracy as the primary metric, assigning 1 for correct and 0 for incorrect responses. Additionally, we employ Gemini 1.5 Pro (Team et al., 2024b) for post-processing model outputs to extract & match the final answer.

Models. LLaVA-NeXT-34B and Phi-4 are two primary backbones, used for modling SpatialMath. To evaluate generalization, we experiment with general, math-, and geometry-focused MSLMs of various sizes; LLaVA-NeXT-7B, LLaVA-OV-7B, InternLM-XC2, Phi-4 (Abdin et al., 2024a), MultiMath-7B (Peng et al., 2024), MATH-LLaVA, G-LLaVA (Gao et al., 2023), and LLaVA-OV-70B (Li et al., 2024).

Baselines. We compare SpatialMath against five strong baselines: (1) Zero-shot: direct model response without additional training. (2) In-context learning (ICL) (Brown et al., 2020): providing 3-shot demonstrations in the prompt. (3) Chain-of-thought (CoT) prompting (Wei et al., 2023):

Models	All		Text Dominant		Text Lite		Vision Intensive		Vision Dominant		Vision Only	
(a): LLaVA-NeXT-34B as Base Model for both SpatialMath-SX **and** SpatialMath-RX **(b): Phi-4 as Base Model for both** SpatialMath-SX **and** SpatialMath-RX												
	(a)	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**
Zero-Shot	18.0	28.2	25.0	34.0	18.0	30.0	21.0	30.0	13.0	27.0	13.0	20.0
ICL (1-shot)	10.0	28.4	15.0	32.0	10.0	33.0	5.0	31.0	11.0	22.0	9.0	24.0
CoT (1-shot)	17.2	28.0	27.0	31.0	18.0	33.0	15.0	30.0	17.0	22.0	9.0	24.0
SFT	18.0	33.4	22.0	38.0	20.0	35.0	22.0	33.0	15.0	30.0	11.0	31.0
SFT + Data Augmentation	19.2	24.6	21.0	35.0	23.0	25.0	18.0	22.0	**20.0**	22.0	14.0	19.0
SpatialMath	**23.0**	**43.6**	27.0	**53.0**	**23.0**	**48.0**	**28.0**	**45.0**	17.0	**41.0**	**20.0**	**31.0**

Table 1: Comparison of SpatialMath with baselines across various problem settings on accuracy scores. one tail Mann-Whitney U test for p-values. We observe one tail Mann-Whitney U test-based p-value of 0.03 and 0.07 for Phi-4 and LLaVA-based SpatialMath when compared to nearest baselines across visual infusions, respectively.

Solver Models	All		Text Dominant		Text Lite		Vision Intensive		Vision Dominant		Vision Only	
(a): Default Solver Performance (Zero-shot Prompting) **(b): Default Solver Performance with LLaVA-NeXT-34B-based** SpatialMath-SX												
	(a)	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**
Phi3.5	22.4	25.4	23.0	25.0	20.0	30.0	23.0	26.0	26.0	22.0	20.0	24.0
InternLMXC2	22.4	23.6	25.0	23.0	26.0	27.0	29.0	25.0	15.0	26.0	17.0	17.0
LLaVA-NeXT-34B	18.0	19.4	25.0	28.0	18.0	18.0	21.0	19.0	13.0	15.0	13.0	17.0
LLaVA OV 7B	30.6	35.4	34.0	43.0	36.0	38.0	34.0	38.0	26.0	33.0	23.0	25.0
LLaVA OV 70B	34.4	35.4	46.0	50.0	36.0	39.0	37.0	36.0	34.0	33.0	19.0	19.0
MultiMath-7B ◇	24.4	25.8	33.0	36.0	28.0	23.0	23.0	29.0	29.0	26.0	9.0	15.0
Math-LLaVA ◇	21.6	23.4	27.0	22.0	21.0	27.0	25.0	32.0	16.0	21.0	19.0	15.0
G-LLaVA *	22.6	23.4	23.0	22.0	23.0	26.0	22.0	26.0	23.0	20.0	22.0	23.0

Table 2: Performance increase across various MSLMs attributed to Spatial Comprehension produced by the LLaVA-NeXT-34B-based SpatialMath-SX (◇ and * represent Math- and domain-focused MSLMs, respectively).

prompting the model to explicitly generate intermediate reasoning steps before answering. (4) Supervised fine-tuning (SFT) (Radford et al., 2021): fine-tuning directly on (q^j, a^j) pairs from the training set. (5) SFT + Data Augmentation: using Spatial Comprehension (SC) and Reasoning Chain (r) to fine-tune a standalone solver, inspired by MultiMath-7B (Peng et al., 2024).

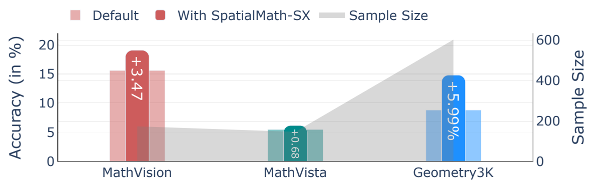


Figure 4: Performance of LLaVA-NeXT-34B-based SpatialMath-SX alogn with default LLaVA-NeXT-34B-based solver across three other public benchmarks- MathVision, MathVista, and Geometry3K.

5 Results

5.1 Performance Comparison

SpatialMath demonstrates substantial performance gains over baseline methods, particularly in vision-dense settings across both base models– LLaVA-NeXT-34B and Phi-4. As shown in Table 1, for LLaVA-NeXT-34B, this approach yields a 10.0 percentage point improvement in vision-intensive settings and a 6.0 percentage point improvement in vision-only settings compared to the Supervised Fine-tuning (SFT) + Data Augmentation baseline. When averaged across all settings, SpatialMath achieves a 3.8 percentage point mean accuracy improvement, demonstrating robustness across formats. For Phi-4, it outperforms SFT baseline with a notable 10.2 percentage point mean accuracy gain across all settings. These findings highlight the strength of infusing Spatial Comprehension into structured Reasoning Chain for addressing visually complex mathematical problems.

5.2 Robustness: SpatialMath-SX

5.2.1 The significance of Spatial Comprehension Core Across MSLMs

As shown in Table 2, passing Spatial Comprehension extracted by SpatialMath-SX as an additional context to default MSLMs yields consistent improvements across various settings. Specifically, we observe accuracy gains up to 7.0, 11.0, and 6.0 percentage points for

Setting	BLEU4	R1	RLSum	Met	BERTScore
SpatialMath-SX: LLaVA-NeXT-34B					
Default Zero-shot	8.8	49.0	45.6	22.9	84.7
SpatialMath-SX	**38.3**	**66.8**	**64.7**	**54.2**	**89.7**
SpatialMath-SX: LLaVA-NeXT-7B					
Default Zero-shot	11.4	47.9	44.9	25.8	84.4
SpatialMath-SX	**14.5**	**51.0**	**49.2**	**41.3**	**87.2**

Table 3: Implicitly evaluate the fine-tuned SpatialMath-SX by comparing generated SC with ground truth using lexical and semantic metrics.

vision-only, vision-dominant, and vision-intensive settings, respectively, leading to an overall mean accuracy improvement of 2.0 (range: 0.8-4.8) percentage points across all settings and solver models. Notably, in the vision-dominant setting, InternLM-XC2 and LLaVA-OV-7B achieve substantial gains of 11.0 and 7.0 percentage points, respectively, while LLaVA-NeXT-34B exhibits a 4.0 percentage point boost in the vision-only setting. These results highlight the robustness and scalability of SpatialMath-SX's structured multimodal comprehension, particularly when applied to specialized visual mathematical models such as MultiMath-7B, reinforcing its efficacy in diverse multimodal reasoning tasks. In response to RQ1, the empirical results indicate that the performance of MSLMs declines as the complexity of visual inputs increases, struggling to understand and process intricate visual information effectively. Further, findings indicate that MSLMs with enriched textual spatial representation and reasoning chains improve visual math problem-solving in response to RQ2.

5.2.2 Evaluation Across Public Benchmarks

To evaluate data generalization, we used three recent public visual-math benchmarks—MathVision (173 samples), MathVista (146), and Geometry3K (601)—which include out-of-distribution samples. To mitigate out-of-domain issues, we focus exclusively on geometry-based problems within these datasets. Figure 4 illustrates the performance difference of default MSLM's with and without the SpatialMath-SX's output as additional context, utilizing the LLaVA-NeXT-34B-based fine-tuned model trained on MathVerse training data. The results demonstrate that MSLM's with Spatial Comprehension yields improvements of 3.47, 0.68, and 5.99 percentage points for MathVision, MathVista, and Geometry3K, respectively, thereby corroborating the robust generalization across other datasets.

5.2.3 Implicit Evaluation

We evaluate fine-tuning robustness by assessing SC quality across SpatialMath-SX models of different sizes, beyond task accuracy. Generated SC are compared to GPT-4o ground-truth using BLEU, ROUGE (R1, RLSum), METEOR, and BERTScore. Table 3 shows fine-tuned SpatialMath-SX consistently outperforming zero-shot models. Mean improvements of 16.3 (BLEU-4), 10.45 (R1), 11.7 (RLSum), 23.4 (METEOR), and 3.9 (BERTScore) confirm fine-tuning enhances SC richness and alignment with ground truth.

5.3 Generalization: SpatialMath-RX

To assess SpatialMath-RX's generalization, we examine if a reasoning solver improves with Spatial Comprehension from different specialized SpatialMath-SX models. Specifically, we investigate two 7B-sized SpatialMath-RX – LLaVA-NeXT-7B and LLaVA-OV-7B, and assess their performance using SC generated by fine-tuned SpatialMath-SX variants. For LLaVA-NeXT-7B, we consider two SpatialMath-SX variants—LLaVA-NeXT-34B and LLaVA-NeXT-7B, while for LLaVA-OV-7B, we use LLaVA-NeXT-34B and LLaVA-OV-7B-based SpatialMath-SX.

Table 4 shows that using SC generated from the larger SpatialMath-SX (LLaVA-NeXT-34B) model improves LLaVA-OV-7B's mean accuracy by 5.0% points across all settings. This confirms that even relatively smaller SpatialMath-RX can benefit from multimodal interpretations generated by a more capable SpatialMath-SX. However, we observe an inconsistent decline in performance when SpatialMath-RX rely on SC from smaller (7B-sized) SpatialMath-SX. Notably, LLaVA-OV-7B exhibits an 8.0 percentage point drop in accuracy when using SC generated from LLaVA-OV-7B itself, highlighting the limitations of smaller models in producing visual interpretations– even after fine-tuning.

5.4 Evaluator Modeling: SpatialMath-SX

Inadequate quality of the SpatialMath-SX's output (SC) may adversely affect the SpatialMath-RX core's performance. To address this, we introduce an intermediate evaluation step for SC between the SpatialMath-SX and SpatialMath-RX cores. To this end, we finetune an evaluative model for a binary decision to determine whether the SC is useful in relation to the underlined question. Based

Solver Models	All		Text Dominant		Text Lite		Vision Intensive		Vision Dominant		Vision Only	
`SpatialMath-RX`: LLaVA-NeXT-7B `SpatialMath-SX`: (a): LLaVA-NeXT-34B (b): LLaVA-NeXT-7B												
	(a)	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**
Zero-shot	18.0	14.2	25.0	17.0	18.0	13.0	21.0	15.0	13.0	15.0	13.0	11.0
`SpatialMath-SX` + default solver	12.4	15.2	12.0	13.0	11.0	17.0	15.0	17.0	13.0	15.0	11.0	14.0
`SpatialMath`	15.4	13.6	14.0	14.0	13.0	13.0	18.0	14.0	18.0	13.0	14.0	14.0
`SpatialMath-RX`: LLaVA-OV-7B `SpatialMath-SX`: (a): LLaVA-NeXT-34B (b): LLaVA-OV-7B												
Zero-shot	30.6	30.6	34.0	34.0	36.0	36.0	34.0	34.0	26.0	26.0	23.0	23.0
`SpatialMath-SX` + default solver	35.6	30.2	44.0	39.0	38.0	32.0	38.0	30.0	33.0	28.0	25.0	22.0
`SpatialMath`	30.0	22.6	40.0	26.0	32.0	24.0	28.0	24.0	25.0	17.0	25.0	22.0

Table 4: Comparison of two distinct `SpatialMath-SX` models, focusing on a diverse range of parameters associated with two `SpatialMath-RX` – LLaVA-NeXT-7B and LLaVA-OV-7B, across a variety of problem settings.

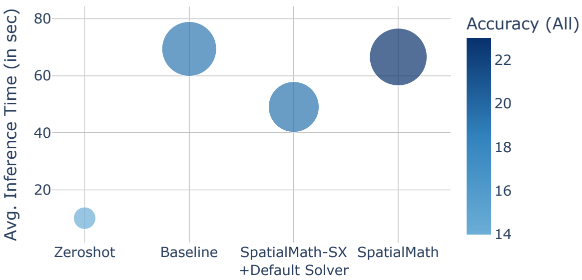


Figure 5: Analysis on computational overhead in comparison to a single model baseline (SFT+Data Aug.) approach. Bubble Size: Average #tokens generated

on the decision, we proceed to pass the `SC` to `SpatialMath-RX` exclusively when the evaluator model issues a "yes" response. To evaluate the impact of this evaluative model, we utilize the improvement ratio, defined as the overall improvement divided by the overall degradation in utilizing `SC` with `SpatialMath-RX` to quantify the effectiveness of quality control. We compare two configurations: `SpatialMath-SX` with and without the evaluative model across both baseline models, LLaVA-NeXT-34B and Phi-4. Notably, evaluative model enhances the improvement ratio from 1.13 to 2.0 for LLaVA-NeXT-34B and from 1.12 to 3.0 for Phi-4, respectively. (refer to Appendix C for more details on evaluator modeling and metric.)

5.5 Computational Efficiency Analysis

The experimental results from Figure 5 demonstrate that `SpatialMath` method is computationally efficient compared to the integrated solver (SFT + data augmentation) approach. While both methods achieve comparable inference time, with `SpatialMath` outperforming on all metrics (Accuracy: 23 vs. 19.2) and lower average inference time (66.59 vs. 69.41 seconds). This highlights the method's effectiveness in balancing computational overhead with improved performance outcomes. Inference times are averaged over three runs on a random set of 100 samples. Additionally, we report the total number of tokens (word-level) generated by the model in each settings. Furthermore, `SpatialMath` is 16× more computationally efficient, requiring only two A100 80GB GPU compared to MathPUMA's 32 A100 80GB GPUs.

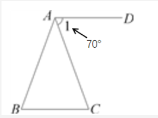
5.6 Error Analysis

Our failure case analysis reveals that a primary source of error stems from inaccurate `SC` generated by the `SpatialMath-SX`, affecting the solver's response quality adversely. Figure 6 highlights an example from a text-dominant setting, where an LLaVA-NeXT-34B-based `SpatialMath-SX` produces an incorrect comprehension of the problem. This erroneous `SC` leads LLaVA-NeXT-34B-based `SpatialMath-RX` to deviate from the correct solution, ultimately generating an incorrect answer. These findings emphasize the importance of quality control mechanisms in refining `SpatialMath-SX` outputs, suggesting potential solutions on error mitigation strategies. Accordingly, we explore an intermediate step to evaluate the quality of `Spatial Comprehension`, generate by `SpatialMath-SX`, prior to forwarding them to the `SpatialMath-RX`. In the future, research community can investigate alternative approaches, such as reinforcement learning from human feedback or cross-context validation to address these challenge.

6 Related Work

Mathematical Reasoning in LLMs. Advancements in MLLMs, such as GPT-4o (OpenAI et al., 2024) and Gemini (Team et al., 2024a), have significantly improved context understanding (Brown et al., 2020) and logical reasoning (Wei et al.,

Original Question
Please directly answer the question and provide the correct option letter, e.g., A, B, C, D.
Question: As shown in the figure, in triangle ABC, AB = AC, passing point A to draw AD parallel BC. If angle 1 = 70.0, then the size of angle BAC is ()
Choices:
A: 40°   B: 30°   C: 70°   D: 50°



SpatialMath-SX Output
Here is an additional context which may help in solving the subsequent question.
Context: ....
4. **Angle 1 = 70: This is an angle within the triangle ABC. It is the angle at vertex A between sides AB and AC.
....

Default Solver	Solver with SpatialMath-SX
• A: 40°	• C: 70°

Reasons for incorrectness of SC
Point 4 states that Angle 1 is angle formed by AB and AC, which is incorrect as Angle 1 is formed due to CA and AD

Figure 6: An example of the unfavorable consequences of misleading spatial comprehension caused by the SpatialMath-SX on the SpatialMath-RX.

2022), enabling applications in mathematical reasoning and word problem-solving. Early research, including MATH (Hendrycks et al., 2021) and MWP-BERT (Liang et al., 2022), focused on text-only problem-solving, excel in algebraic manipulation and numerical inference but struggle with spatial reasoning (Lu et al., 2023).

Visual Mathematical Reasoning. Models such as LLaVA (Liu et al., 2023), InternLM-XC (Zhang et al., 2023), Phi-3 (Abdin et al., 2024b), and Pixtral (Agrawal et al., 2024) have shown strong performance in visual understanding tasks such as image captioning and VQA (Agrawal et al., 2016), demonstrating impressive accuracy (Liu et al., 2024a) but require additional techniques to align symbolic notations with their visual representations when applied in mathematical contexts (Yan et al., 2024). Dedicated efforts to integrate symbolic reasoning with diagrammatic understanding have led to specialized multimodal models, including MATH-Vision (Wang et al., 2024), MATH-VISTA (Lu et al., 2024), Math-LLaVA (Shi et al., 2024b), and CLEVR-Math (Lindström and Abraham, 2022). They aim to preserve the syntactic integrity of mathematical expressions while incorporating spatial reasoning from diagrams.

Novelty of Our Work. MathVerse (Zhang et al., 2024a) analyzes visual mathematical reasoning by assessing model performance with different levels of visual infusion. However, it lacks a systematic method for improving performance in vision-dense contexts, a critical gap that our research effectively addresses. Early approaches, such as MAVIS (Zhang et al., 2024b) and Math-PUMA (Zhuang et al., 2024), utilized multi-stage (3-4) training pipelines with large-scale synthetic data, but faced challenges in computational efficiency and scalability to broader mathematical domains. These methods often rely heavily on synthetic data generation, limiting practical applicability. Thus, prior efforts remain computationally expensive or exhibit inconsistent performance in dense visual mathematical reasoning, particularly due to tackling these tasks in isolation. SpatialMath addresses these challenges by advancing visual comprehension-infused symbolic reasoning in a computationally efficient framework.

7 Conclusion

In this paper, we introduced a novel framework, SpatialMath, supplemented by a new dataset, MATHVERSE-PLUS. This approach aims to enhance visual spatial comprehension and infuse it into the sound symbolic reasoning capabilities of MSLMs across a diverse range of visual and textual integrations, specifically addressing geometric mathematics problems. Empirical results reinforce that SpatialMath surpasses contemporary baselines, thereby representing a significant advance in developing medium-sized visual mathematics solvers, which can proficiently interpret complex visual mathematical problems and make sound logical deductions.

8 Limitations

The proposed dataset, MATHVERSE-PLUS is derived from an existing dataset, which necessitates reliance on the authors of the source dataset to establish the quality standards applied. Further, the efficacy of the reasoning-enhanced solver, SpatialMath-RX, is contingent upon the outcomes generated by the SpatialMath-SX. Therefore, any errors originating from the SpatialMath-SX: Spatial Comprehension Core may propagate to the SpatialMath-RX, potentially making an adverse impact, despite the mitigation strategies presented in this work.

Additionally, it is important to note that the dataset is exclusively based on the English language, thereby limiting our ability to assess the applicability and effectiveness of the proposed methodology in non-English contexts, particularly for low-resource languages.

9 Ethical Considerations

The proposed dataset, referred to as MATHVERSE-PLUS represents an expansion of an existing publicly available dataset and will be released publicly in accordance with the source dataset's guidelines. All data annotators involved in this study were compensated fairly for their contributions. We recognize the lack of women's representation in MATHVERSE-PLUS annotations as a reflection of the broader global gender disparity in research and innovation; however, the authors do not endorse this perspective. We are committed to promoting gender and racial equality within research fields.

Furthermore, we clarify that no generative AI tools were utilized in the creation of this content, aside from those used for spell-checking and grammar correction. We emphasize that the scope of MATHVERSE-PLUS is strictly dedicated to scientific research purposes. The selection of GPT-4o for data augmentation was a decision made based on its popularity, and the authors do not express bias toward any specific commercial large language model.

Acknowledgments

Tanmoy Chakraborty acknowledges the support of Microsoft Research Grant, Azure AI Credits Grant from Microsoft's Accelerating Foundation Models Research (AFMR) initiative, Google GCP Grant, and Rajiv Khemani Young Faculty Chair Professorship in Artificial Intelligence.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024a. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Marah Abdin, Jyoti Aneja, and Others. 2024b. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2016. Vqa: Visual question answering. *Preprint*, arXiv:1505.00468.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, and 23 others. 2024. Pixtral 12b. *Preprint*, arXiv:2410.07073.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. 2021. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523, Online. Association for Computational Linguistics.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, and 4 others. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *Preprint*, arXiv:2401.16420.

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

Yunhao Ge, Xiaohui Zeng, Jacob Samuel Huffman, Tsung-Yi Lin, Ming-Yu Liu, and Yin Cui. 2024. Visual fact checker: Enabling high-fidelity detailed caption generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Preprint*, arXiv:2403.14608.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

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. *Preprint*, arXiv:2106.09685.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Zhenwen Liang, Jipeng Zhang, Lei Wang, Wei Qin, Yunshi Lan, Jie Shao, and Xiangliang Zhang. 2022. MWP-BERT: Numeracy-augmented pre-training for math word problem solving. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 997–1009, Seattle, United States. Association for Computational Linguistics.

R. Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22 140:55–55.

Adam Dahlgren Lindström and Savitha Sam Abraham. 2022. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *Preprint*, arXiv:2208.05358.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6774–6786, Online. Association for Computational Linguistics.

Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. 2023. A survey of deep learning for mathematical reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14605–14631, Toronto, Canada. Association for Computational Linguistics.

OpenAI, Josh Achiam, and Others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2024a. Continual learning of large language models: A comprehensive survey. *Preprint*, arXiv:2404.16789.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. 2024b. Math-LLaVA: Bootstrapping mathematical reasoning for multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4663–4680, Miami, Florida, USA. Association for Computational Linguistics.

Gemini Team, Rohan Anil, and Others. 2024a. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Gemini Team, Petko Georgiev, and Others. 2024b. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. *Preprint*, arXiv:1411.4555.

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, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Yibo Yan, Jiamin Su, Jianxiang He, Fangteng Fu, Xu Zheng, Yuanhuiyi Lyu, Kun Wang, Shen Wang, Qingsong Wen, and Xuming Hu. 2024. A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges. *Preprint*, arXiv:2412.11936.

Pan Zhang, Xiaoyi Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, Xinyu Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Y. Qiao, Da Lin, and Jiaqi Wang. 2023. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *ArXiv*, abs/2309.15112.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. 2024a. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *Preprint*, arXiv:2403.14624.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, Peng Gao, Chunyuan Li, and Hongsheng Li. 2024b. Mavis: Mathematical visual instruction tuning with an automatic data engine. *Preprint*, arXiv:2407.08739.

Wenwen Zhuang, Xin Huang, Xiantao Zhang, and Jin Zeng. 2024. Math-puma: Progressive upward multimodal alignment to enhance mathematical reasoning. *Preprint*, arXiv:2408.08640.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. 2025. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *Preprint*, arXiv:2411.00836.

A Data Statistics and Quality

A.1 MathVerse.

MathVerse provides a diverse range of visual infusion in geometric mathematical problems comprising multi-choice and free-form style Q/A pairs. It serves as a contemporary benchmark for evaluating MSLMs with respect to their visual mathematical reasoning capabilities. The dataset was derived from various publicly accessible datasets and subsequently transformed to incorporate varied degrees of visual infusion in mathematical inquiries.

The construction of the questions within MathVerse posits that a multiple-choice geometry problem typically comprises three components in addition to the answer choices – descriptive information, an implicit property, and an essential condition. Descriptive information pertains to the directly observable and distinctly represented elements within the diagram, such as the presence of geometric shapes or the intersection points of functions. The implicit property necessitates a higher level of visual perception to infer from the diagram, signifying critical visual conditions for problem-solving. This includes aspects such as the parallelism and perpendicularity of lines, the similarity and congruence of triangles, as well as the classification and periodicity of functions. The essential condition represents specific numerical or algebraic measurements. By leveraging these three properties along with the accompanying diagram, MathVerse presents six versions of each problem – Text-Dominant, Text-Only, Text-Lite, Vision-Intensive, Vision-Dominant, and Vision-Only. Notably, the Vision-Only configuration presents the mathematical problem solely in visual modality, devoid of any textual information.

A.2 `MATHVERSE-PLUS` Statistics.

In the construction of the `MATHVERSE-PLUS`, this study focuses exclusively on 2D and 3D geometric problems, deliberately excluding issues related to functions to ensure the dataset is homogenized for geometry-related purposes. Additionally, we consider only five visual modalities derived from the source dataset: Text-Dominant, Text-Lite, Vision-Intensive, Vision-Dominant, and Vision-Only. The Text-Only modality has been excluded, as it does not align with the multimodal focus of this research. During the fine-tuning phases, a total of 2,260 instances were considered across all settings, while 500 instances were allocated for testing purposes in the results presented herein.

To ensure the quality of the augmented data, we exclusively consider the `Reasoning Chain` (r) in which the model's final answer aligns with the ground truth during the generation process. This approach resulted in the selection of 1,690 high-quality training instances for the `SpatialMath-RX` and the (SFT + data augmentation) baseline fine-tuning. In order to facilitate the model's learning of a consistent `Reasoning Chain`, we utilize the `Reasoning Chain` (r) derived from the Text-Dominant setting across all ranges of visual infusion. This strategy provides the model with the maximum input necessary for the generation of the `Reasoning Chain`.

A.3 Commentaries on Commercial LLM's Usage.

The recent advancements in open-source LLMs, exemplified by DeepSeek-R1 with publicly available code and weights, demonstrate performance comparable to that of commercial LLMs. This development significantly alleviates the accessibility challenges posed by proprietary systems such as GPT-

4. The data augmentation capabilities of GPT-4 can be effectively substituted with the open-source DeepSeek-V3 (DeepSeek-AI, 2024) or LLaMA-4 [2]-Maverick.

In Figure 7, we present a qualitative analysis of GPT4o and LLaMA-4-Maverick in generating `Spatial Comprehension`. Analysis reveals that both models demonstrate an impressive ability to interpret the visual math problem and translate it into a structured, conceptual breakdown. They effectively identify the core geometric components, such as the triangle's given angles and the crucial relationship created by the parallel lines. The models showcase two distinct yet valuable descriptive styles. GPT4o excels at providing a comprehensive overview of the geometric properties, correctly identifying concepts like similar triangles and corresponding angles, which lays a robust theoretical foundation. On the other hand, LLaMA-4-Maverick adopts a more direct, pedagogical approach, outlining a clear, step-by-step path to the solution by highlighting the key principles needed.

B Extended Results

B.1 Performance Across Benchmarks

In the experiment, the selected categories of data for Geometry 3K, MathVista, and MathVision were carefully defined to ensure thorough Geometry-based visual math analysis. For Geometry 3K, the entire test set comprised 610 samples, allowing for a comprehensive exploration of geometry-related concepts. In the case of MathVista, the data focused specifically on metrics associated with angle measurement within the subject of metric geometry - angle. Meanwhile, MathVision was centered around the task of geometry problem-solving, emphasizing language proficiency in English (geometry problem solving and language - english). The detailed results are presented in Table 5.

B.2 Why `SpatialMath-SX`: Spatial Comprehension Core Is Important

In this experiment, we evaluate the effectiveness of incorporating `Spatial Comprehension` into structured reasoning chains using evolutionary analysis. First, we fine-tune an additional `SpatialMath-SX` variant, referred to as `SpatialMath-SX`$_{text\_only}$, which utilizes the text from 'Text Only' setting, the maximal textual description available for a given question as an alternative `SC`. This formulation enables a systematic assessment of the impact of providing detailed available textual problem descriptions across diverse visual contexts. We compare performance across three settings: the default solver with `SpatialMath-SX`$_{text\_only}$, the default solver with `SpatialMath-SX`, and the `SpatialMath`, applied to two base architectures, LLaVA-NeXT-34B and Phi-4. The remaining experimental configurations align with those outlined in the primary study within the paper.

The findings presented in Table 6 demonstrate two key observations. First, the inclusion of a detailed `Spatial Comprehension`, which encapsulates geometric elements and relationships, significantly improves the performance of the default solver when compared to utilizing only the maximal textual description of the problem. Second, the integration of `Spatial Comprehension` into `Reasoning Chain` yields substantial enhancements across models and diverse visual-textual settings. These results underscore the effectiveness of the novel proposed methodology, visual-comprehension-aware structured symbolic reasoning (`SpatialMath`), in augmenting the problem-solving capabilities of MSLMs.

B.3 Computational Efficiency Analysis

The detailed experimental results are presented in Table 7.

B.4 Robustness Across Temperature Parameter

The section demonstrates the robustness of the proposed methodology through the manipulation of the model parameter known as Temperature, which governs the randomness of the output. The findings presented in Table 8 elucidate that the proposed method exhibits a high degree of robustness across various probabilistic vocabulary spaces, as influenced by the parameter temperature across three runs.

B.5 Commentaries on Closed-Source Models

Table 9 presents the evaluation results for the MATHVERSE benchmark for closed-source models. These results are sourced from the original publication by (Zhang et al., 2024a) and are provided here purely for reference purposes. We observe that the exponential increase in the num-

[2] https://ai.meta.com/blog/llama-4-multimodal-intelligence

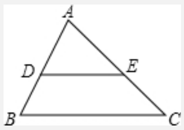
Setting: Text Dominant
Prompt:

- Please generate the additional context for above question which may help in better comprehension of the problem. Context may include elaboration and simplification of the problem such that it becomes easier to solve it. You must not provide either the steps involve in the solution or the final solution to the problem

Question:

- As shown in the figure, in triangle ABC, it is known that angle A = 80.0, angle B = 60.0, point D is on AB and point E is on AC, DE parallel BC, then the size of angle CED is () Choices: A: 40° B: 60° C: 120° D: 140°

Image:



Response - GPT4o:

- To better understand the problem, let's break down the given information and the relationships within the triangle: 1. **Triangle ABC**: This is the main triangle in the problem. We are given two angles: *Angle A = 80° *Angle B = 60°

  2. **Point D on AB and Point E on AC**: These points divide the sides AB and AC respectively.

  3. **DE is parallel to BC**: This is a crucial piece of information. When a line segment (DE) is parallel to one side of a triangle (BC) and intersects the other two sides (AB and AC), it creates similar triangles. In this case, triangle ADE is similar to triangle ABC.

  4. **Angle Relationships**:

  * Since DE is parallel to BC, angle ADE is equal to angle ABC, and angle AED is equal to angle ACB due to the properties of parallel lines and corresponding angles.

  5. **Finding Angle CED**:

  * To find angle CED, we need to use the properties of similar triangles and the given angles. Since DE is parallel to BC, angle CED will have a specific relationship with the angles in triangle ABC.

  By understanding these relationships and properties, you can determine the size of angle CED using the given angles and the fact that DE is parallel to BC. This context should help in comprehending the problem and setting up the necessary steps to find the solution.

Response - LLaMA-4-Maverick:

- **Problem Context and Elaboration**

  The given problem involves a triangle ABC with specific angle measures and points D and E on sides AB and AC, respectively. The task is to determine the size of angle CED given that DE is parallel to BC.

  Given Information: * Triangle ABC with known angles: angle A = 80.0 degrees and angle B = 60.0 degrees. * Point D lies on side AB. * Point E lies on side AC. * DE is parallel to BC.

  Objective: To find the measure of angle CED.

  Relevant Geometric Principles: * The sum of the interior angles of a triangle is always 180 degrees. * When two lines are parallel, corresponding angles are equal.

  Understanding the Figure: Since DE is parallel to BC, corresponding angles formed by the transversal through points A, E, and C will be equal. This implies that angle AED is equal to angle ACB due to the corresponding angles property.

  Calculating Angle ACB: Using the fact that the sum of the interior angles of a triangle is 180 degrees, we can calculate angle ACB as follows:

  angle A + angle B + angle C = 180 degrees 80 degrees + 60 degrees + angle C = 180 degrees angle C = 180 degrees - 140 degrees angle C = 40 degrees

  Implication for Angle CED:

  Since DE is parallel to BC, angle AED = angle ACB = 40 degrees. Knowing this, we can use the property that the sum of the angles on a straight line is 180 degrees to find angle CED. angle AED + angle CED = 180 degrees Substituting the known value of angle AED: 40 degrees + angle CED = 180 degrees This sets the stage for calculating angle CED.

  The problem requires using the properties of parallel lines and the angles within a triangle to solve for angle CED.

Figure 7: A qualitative analysis of `Spatial Comprehension` generational capabilities of GPT4o and LLaMA-4.

Benchmarks	Default	With SpatialMath-SX	Improvements(in %)	Samples Size
MathVision	15.61	19.08	3.47	173
MathVista	5.48	6.16	0.68	146
Geometry3K	8.82	14.81	5.99	601

Table 5: (Detail results) Performance of LLaVA-NeXT-34B-based SpatialMath-SX alogn with default LLaVA-NeXT-34B-based solver across three other benchmarks- MathVision, MathVista, and Geometry3K.

Models	All		Text Dominant		Text Lite		Vision Intensive		Vision Dominant		Vision Only	
(a): LLaVA-NeXT-34B as Base Model for both SpatialMath-SX and SpatialMath-RX **(b): Phi-4 as Base Model for both SpatialMath-SX and SpatialMath-RX**												
	(a)	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**
SpatialMath-SX$_{text_only}$	18.0	29.2	25.0	35.0	18.0	30.0	21.0	30.0	13.0	30.0	13.0	21.0
SpatialMath-SX	19.4	30.4	**28.0**	36.0	18.0	34.0	19.0	31.0	15.0	26.0	17.0	25.0
SpatialMath	**23.0**	**43.6**	27.0	**53.0**	**23.0**	**48.0**	**28.0**	**45.0**	17.0	**41.0**	**20.0**	**31.0**

Table 6: Evolutionary Analysis for assessing the impact of SpatialMath-SX.

Setting	Inference Time (in Sec)				#Tokens Generated	Accuracy (All)
	run1	run2	run3	Average		
Zeroshot	10.17	9.94	10.20	10.10	73.2	14
SFT + Data Augmentation	68.27	69.45	70.51	69.41	438.05	19.2
SpatialMath-SX	49.41	49.01	48.81	49.08	376.95	19.4
SpatialMath	67.02	66.66	66.10	66.59	484.73	23

Table 7: (Detail Results) Analysis on computational overhead in comparison to a single model approach. Bubble Size: Average #tokens generated.

Solver Models	All	Text Dominant	Text Lite	Vision Intensive	Vision Dominant	Vision Only
SFT + Data Augmentation	19.2	21.0	23.0	18.0	20.0	14.0
Temp = 0	23.0	27.0	23.0	28.0	17.0	20.0
Temp = 0.2	23.0	28.0	23.0	28.0	16.0	20.0
Temp = 1	23.2	29.0	23.0	28.0	17.0	19.0
Average	23.1	28.0	23.0	28.0	16.7	19.7

Table 8: Significance of model's parameter Temperature. Rest of the settings are same as Table 1.

ber of parameters in closed-source models enables them to outperform their open-source counterparts. Despite achieving the highest overall average performance of 63.1%, GPT-4V remains suboptimal, highlighting the need for further efforts to enhance the performance of closed-source models. Notably, advancements in parameter fine-tuning methodologies are typically reliant on access to the model's architecture. However, closed-source models restrict direct access to their architectures, limiting the application of such methodologies to open-source models. Given access to the architectural parameters, there is intense anticipation that strategies akin to SpatialMath could similarly augment the capabilities of closed-source models.. It is important to note that the results encompass all six settings, including samples from geometry-specific problems as well as function-related samples.

C Evaluator Modeling on SpatialMath-SX Cont.

C.1 Objective

The output of the SpatialMath-SX module is contingent upon the quality and geometrical accuracy established within the visual figure generated by the process. Inadequate quality of the Spatial Comprehension (SC) may adversely affect the SpatialMath-RX core's performance. Hence, the primary aim of the Evaluator module is to systematically assess the quality of the generated SC and to filter out any instances deemed unsatisfactory, thereby allowing only those that meet quality standards to proceed to the SpatialMath-RX core. This process is intended to mitigate the detrimental effects of the SpatialMath-SX module on the SpatialMath-RX module. The Evaluator is positioned as an intermediary between

Model	All		Text Dominant		Text Lite		Text Only		Vision Intensive		Vision Dominant		Vision Only	
	CoT-E	Acc	CoT-E	Acc	CoT-E	Acc	CoT-E	Acc	CoT-E	Acc	CoT-E	Acc	CoT-E	Acc
LLMs														
ChatGPT	-	-	51.3	33.3	38.5	18.9	51.3	33.3	-	-	-	-	-	-
GPT-4	-	-	63.4	46.5	40.7	20.7	63.4	46.5	-	-	-	-	-	-
Closed-source MLLMs														
Qwen-VL-Plus	21.3	11.8	26.0	15.7	21.2	11.1	25.2	14.5	18.5	9.0	19.1	13.0	21.8	10.0
Gemini-Pro	35.3	23.5	39.8	26.3	34.7	23.5	44.5	27.3	32.0	23.0	36.8	22.3	33.3	22.2
Qwen-VL-Max	37.2	25.3	42.8	30.7	37.7	26.1	47.9	28.9	33.6	24.1	35.9	24.1	35.9	21.4
GPT-4V	54.4	39.4	63.1	54.7	56.6	41.4	60.3	48.7	51.4	34.9	50.8	34.4	50.3	31.6

Table 9: The below table is sourced from the original work presented by (Zhang et al., 2024a) for the reference purpose only. Mathematical evaluation on six problem versions in MATHVERSE's *testmini* set. The highest accuracy for closed-source MLLMs is marked in red.

the SpatialMath-SX and SpatialMath-RX modules. Various criteria may be interpreted to distinguish a good SC from a bad one. In this context, we define the efficacy of SC based on its utility within the SpatialMath-RX core. A satisfactory Spatial Comprehension is characterized by its lack of adverse effects on the outputs generated by the SpatialMath-RX module.

C.2 Training Data

To achieve this objective, we first develop a training dataset to construct a robust evaluator model. Utilizing this training set of MATHVERSE-PLUS, we create instances where the input comprises a question $q_{(t,v)}$ alongside its corresponding SC produced by the SpatialMath-SX module, with an appended inquiry: "Is the provided context faithful to the given context?" The expected output is a binary label of "yes" or "no," indicating whether the SC has a detrimental impact on the output of the SpatialMath-RX core. A label of 'no' denotes an adverse impact, while 'yes' signifies the absence of such an impact.

C.3 Fine-tuning and Metrics

Subsequently, using the aforementioned training data, we proceed to fine-tune the evaluator models associated with both the base models LLaVA-NeXT-34B and Phi-4, respectively. The configurations utilized in this fine-tuning process are consistent with those employed during the fine-tuning of the SpatialMath-SX and SpatialMath-RX modules. To quantify the impact of the evaluative model, we introduce the concept of the improvement ratio, presented in Equation 5, which is defined as the overall improvement divided by the overall degradation associated with the utilization of SC in conjunction with SpatialMath-RX for the purpose of assessing the effectiveness of quality control. In this context, an improvement refers to the positive changes observed in the outcomes of the SpatialMath-RX module resulting from the application of SC while degradation signifies the adverse effects on the outcomes of the SpatialMath-RX module when employing SC as compared to the default settings.

$$\text{Improvement Ratio} = \frac{\text{Overall Improvements}}{\text{Overall Degradations}} \tag{5}$$

C.4 Results

The experimental results presented in Table 10 highlight the comparative performance of SpatialMath-SX and SpatialMath-SX with Eval model across two base models, LLaVA-NeXT-34B and Phi-4. Both approaches demonstrate noticeable improvements in "Accuracy (All) Improvements" over their respective baseline models, with Phi-4 showing slightly higher enhancement (+2.2% versus +1.4%). When analyzing sample-level metrics, SpatialMath-SX with Eval yields fewer "Absolute Degradations" compared to SpatialMath-SX alone, reflecting the refinement introduced by the evaluator. Notably, the improvement ratio is more pronounced for SpatialMath-SX with Eval, particularly for Phi-4 (3.0 vs. 1.12), suggesting that the evaluator effectively balances improvements and degradations at the sample level. The utility tag further supports this observation, indicating higher 'yes' predictions with SpatialMath-SX with Eval, underscoring its practical efficacy.

Findings also reveal a slight reduction in "Accuracy (All) Improvements" when incorporating the evaluator model, with LLaVA-NeXT-34B and

Phi-4 showing decreases of 0.6% and 1.4%, respectively. This suggests that while the evaluator successfully minimizes "Absolute Degradations" and improves the overall imrpovement ratio, it comes at the cost of a decreased top-level accuracy. These findings highlight the need to optimize between maximizing accuracy and minimizing adverse impacts, striking a balance to ensure both reliability and fairness in model predictions.

D Experimental Setup Cont.

We conduct experiments using LLaVA- and Phi-based architectures to evaluate the proposed method. To mitigate catastrophic forgetting (Shi et al., 2024a) due to limited training data, we freeze the LLM and vision tower components, fine-tuning only the MLP projector for both `SpatialMath-SX` and `SpatialMath-RX` for LLaVA-based model. Similarly, for Phi-4, we fine-tune lora adapter and vision encoder. All experiments are conducted once on same default random seed settings.

D.1 Hyperparamters

In this section, we details all the hyperparameters to ensure reproducibility of the results. LLaVa-based models consist of three primary components: a large language model (LLM) backbone and a vision tower followed by a multilayer perceptron (MLP) adapter coneecting the two. In our approach, we exclusively finetune the MLP adapter while keeping the other two components fixed. The fine-tuning process is conducted over three epochs across various methods. Our computational infrastructure is comprised of a dual NVIDIA A100[3] GPU setup. The detailed hyperparameters for LLaVA-NeXT-34B and Phi-4 finetuning are presented in Table 12 and Table 11, respectively. The peak GPU usage for training either core of the `SpatialMath` was approximately 96 GPU hours, utilizing a shared compute environment.

D.2 Prompts

In this study, we delineate the diverse prompts employed across various settings and fine-tuning procedures. The prompts utilized to evaluate zero-shot, in-context (one-shot), and chain-of-thought (one-shot) methodologies are depicted in Figures 8, 9, and 10, respectively. Additionally, a training data instance that encompasses an input and output for the objectives `SpatialMath-SX`$_{text\_only}$, `SpatialMath-SX` `SpatialMath`, and the SFT+ data augmentation baseline is illustrated in Figures 13, 14, 15, and 16, respectively. Furthermore, the prompts deployed during the inference stages for the `SpatialMath-SX`$_{text\_only}$ and `SpatialMath-SX` fine-tuned study companion models are presented in Figures 17 and 18, respectively. Concurrently, the prompts applied during the inference for the `SpatialMath-RX` are shown in Figure 19. The prompts utilized during the data augmentation phase to generate `SC` and reasoning chain r are illustrated in Figure 11 and Figure 12, respectively.

Furthermore, we clarify that no generative AI tools were utilized in the creation of this content, aside from those used for spell-checking and grammar correction. We emphasize that the scope of `MATHVERSE-PLUS` is strictly dedicated to scientific research purposes. The selection of GPT-4o for data augmentation was a decision made based on its popularity, and the authors do not express bias toward any specific commercial large language model.

[3] https://www.nvidia.com/en-in/data-center/a100/

Metric	Method	LLaVA-NeXT-34B	Phi-4
Accuracy (All) Improvements	`SpatialMath-SX` `SpatialMath-SX` + with Eval	1.4 0.8	2.2 0.8
Absolute Improvements	`SpatialMath-SX` `SpatialMath-SX` + with Eval	61 8	57 6
Absolute Degradations	`SpatialMath-SX` `SpatialMath-SX` + with Eval	54 4	51 2
Improvement Ratio	`SpatialMath-SX` `SpatialMath-SX` + with Eval	1.13 2.0	1.12 3.0
Utility Tag	Yes/500	49	21

Table 10: The comparison of `SpatialMath-SX` with `SpatialMath-SX` using Evaluator across different base models is presented here. "Accuracy (All) improvements" measures the enhancement over the default model, while "Absolute improvements" and "Absolute degradations" are assessed at the sample level. The utility tag indicates the 'yes' predictions made by the evaluator for the test set.

Please directly answer the question and provide the correct option letter, e.g., A, B, C, D.

Question: As shown in the figure, in triangle ABC, it is known that angle A = 80.0, angle B = 60.0, point D is on AB and point E is on AC, DE parallel BC, then the size of angle CED is ()

Choices:
A: 40°
B: 60°
C: 120°
D: 140°

Figure 8: An example of prompt for Zero-shot inference across solver models for default performance benchmark along with an example problem.

Please directly answer the question and provide the final value, e.g., 1, 2.5, 300.

Example 1:
Question: As shown in the figure, Given that AB = 2 DE, angle E = 16.0, then the degree of angle ABC is ()
Choices:
A: 40°
B: 60°
C: 120°
D: 140° <image>
Answer: B

Example 2:
Question: As shown in the figure, in triangle ABC, it is known that angle A = 80.0, angle B = 60.0, point D is on AB and point E is on AC, DE parallel BC, then the size of angle CED is ()
Choices:
A: 40°
B: 60°
C: 120°
D: 140° <image>
Answer:

Figure 9: An example of prompt for one-shot ICL inference across solver models for default performance benchmark along with an example problem.

Please first conduct reasoning, and then answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.

Question: As shown in the figure, it is known that angle A = 80.0, angle B = 60.0, DE parallel BC, then the size of angle CED is ()

Choices:
A: 40°
B: 60°
C: 120°
D: 140°

Figure 10: Prompt for one-shot CoT inference across solver models for default performance benchmark along with an example problem.

As shown in the figure, it is known that angle A = 80.0, angle B = 60.0, DE parallel BC, then the size of angle CED is ()
Choices:

- A: 40°
- B: 60°
- C: 120°
- D: 140°

Please generate the additional context for above question which may help in better comprehension of the problem. Context may include elaboration and simplification of the problem such that it becomes easier to solve it. You must not provide either the steps involve in the solution or the final solution to the problem.

Figure 11: Prompt to generate perceptual interpretation from relatively larger model (GPT-4o): Data augmentation. The problem image was provided along with this prompt as inputs to the model.

Please first conduct reasoning, and then answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.\nQuestion: As shown in the figure, it is known that angle A = 80.0, angle B = 60.0, DE parallel BC, then the size of angle CED is ()
Choices:

- A: 40°
- B: 60°
- C: 120°
- D: 140°

Figure 12: Prompt to generate reasoning chains from relatively larger model (GPT-4): Data augmentation. The problem image was provided along with this prompt as inputs to the model.

A geometry problem often includes three components besides the answer choices: descriptive information, an implicit property, and an essential condition.
Descriptive Information refers to the directly observable and clearly portrayed content in the diagram, such as the presence of geometric shapes or intersection points of functions.
The second part, implicit property, involves a higher level of visual perception to discern from the diagram. It signifies strong visual conditions for problem-solving, such as the parallelism and perpendicularity between lines, the similarity and congruence among triangles, and the category and periodicity of functions.
The third part, essential condition, denotes the specific numerical or algebraic measurements.
Here is an incomplete textual question along with an image representing an associated diagram of a geometry problem where the implicit property and essential condition are given in the text. Can you complete and write this question accurately in text-only form, taking into account the descriptive information presented through the diagram?
Incomplete Textual Question: As shown in the figure, it is known that angle A = 80.0, angle B = 60.0, DE parallel BC, then the size of angle CED is ()
Choices:
A: 40°
B: 60°
C: 120°
D: 140°
Associated Diagram: <image>

As shown in the figure, in triangle ABC, it is known that angle A = 80.0, angle B = 60.0, point D is on AB and point E is on AC, DE parallel BC, then the size of angle CED is ()

Choices:

- A: 40°
- B: 60°
- C: 120°
- D: 140°

Figure 13: An instance of input (highlighted in magenta) and Output for fine-tuning objective of `SpatialMath-RX`.

Please directly answer the question and provide the correct option letter, e.g., A, B, C, D.

Question: As shown in the figure, in triangle ABC, it is known that angle A = 80.0, angle B = 60.0, point D is on AB and point E is on AC, DE parallel BC, then the size of angle CED is ()
Choices:
A: 40°
B: 60°
C: 120°
D: 140° <image>

Based on the information provided, here are a few clues to solve the question:
To better understand the problem, let's break down the given information and the relationships within the triangle:

1. Triangle ABC: This is the main triangle in the problem. We are given two angles:
    - Angle A = 80°
    - Angle B = 60°
2. Point D on AB and Point E on AC: These points divide the sides AB and AC, respectively.
3. DE is parallel to BC: This is a crucial piece of information. When a line segment (DE) is parallel to one side of a triangle (BC) and intersects the other two sides (AB and AC), it creates similar triangles. In this case, triangle ADE is similar to triangle ABC.
4. Angle Relationships:
    - Since DE is parallel to BC, angle ADE is equal to angle ABC, and angle AED is equal to angle ACB due to the properties of parallel lines and corresponding angles.
5. Finding Angle CED:
    - To find angle CED, we need to use the properties of similar triangles and the given angles. Since DE is parallel to BC, angle CED will have a specific relationship with the angles in triangle ABC.

By understanding these relationships and properties, you can determine the size of angle CED using the given angles and the fact that DE is parallel to BC. This context should help in comprehending the problem and setting up the necessary steps to find the solution.

Perceptual Interpretation of the Question:

As shown in the figure, in triangle ABC, it is known that:

- Angle A = 80.0°
- Angle B = 60.0°
- Point D is on AB and point E is on AC
- DE is parallel to BC

Then, the size of angle CED is:

Choices:

- A: 40°
- B: 60°
- C: 120°
- D: 140°

Figure 14: An instance of input (highlighted in magenta) and Output for fine-tuning objective of `SpatialMath-SX`.

Carefully follow the instructions to solve the following geometry question.\n\nInstructions:
1. Simplify the question: Break down the problem so it becomes easier to understand and solve.
2. 2. Reconstruct the question: After simplification, rewrite the question in a clear, text-only format.
3. 3. Apply stepwise reasoning: Solve the problem logically by taking each step methodically.
4. 4. Report the final answer: Conclude the solution with the final, correct answer\n\nHere is the question:
As shown in the figure, in triangle ABC, it is known that angle A = 80.0, angle B = 60.0, point D is on AB and point E is on AC, DE parallel BC, then the size of angle CED is ()
Choices:
A: 40°
B: 60°
C: 120°
D: 140° <image>

To better understand the problem, let's break down the given information and the relationships within the triangle:

1. Triangle ABC: This is the main triangle in the problem. We are given two angles:
   - Angle A = 80°
   - Angle B = 60°
2. Point D on AB and Point E on AC: These points divide the sides AB and AC, respectively.
3. DE is parallel to BC: This is a crucial piece of information. When a line segment (DE) is parallel to one side of a triangle (BC) and intersects the other two sides (AB and AC), it creates similar triangles. In this case, triangle ADE is similar to triangle ABC.
4. Angle Relationships:
   - Since DE is parallel to BC, angle ADE is equal to angle ABC, and angle AED is equal to angle ACB due to the properties of parallel lines and corresponding angles.
5. Finding Angle CED:
   - To find angle CED, we need to use the properties of similar triangles and the given angles. Since DE is parallel to BC, angle CED will have a specific relationship with the angles in triangle ABC.

By understanding these relationships and properties, you can determine the size of angle CED using the given angles and the fact that DE is parallel to BC. This context should help in comprehending the problem and setting up the necessary steps to find the solution.

Hence the simplified question becomes as follows:
As shown in the figure, in triangle ABC, it is known that angle A = 80.0°, angle B = 60.0°, point D is on AB, and point E is on AC. Since DE is parallel to BC, the size of angle CED is: ().

Let's solve the question step-by-step:

Since DE is parallel to BC, triangle ADE is similar to triangle ABC by the Basic Proportionality Theorem (also known as Thales' theorem). This means that the corresponding angles of these triangles are equal.

Given:

- Angle A = 80°
- Angle B = 60°

In triangle ABC:

- Angle C = 180° - Angle A - Angle B = 180° - 80° - 60° = 40°

Since DE is parallel to BC:

- Angle ADE = Angle ABC = 60°
- Angle AED = Angle ACB = 40°

In triangle ADE:

- Angle DAE = Angle A = 80°
- Angle ADE = 60°
- Angle AED = 40°

Now, angle CED is an exterior angle for triangle ADE at vertex E. The exterior angle is equal to the sum of the two non-adjacent interior angles.

Thus:

- Angle CED = Angle ADE + Angle DAE = 60° + 80° = 140°

The correct option is:

- D: 140°

Therefore, the correct answer is D.

Figure 15: An instance of input (highlighted in magenta) and Output for fine-tuning objective of (SFT+Data Augmentation) baseline.

Here is an additional context which may help in solving the subsequent question.

Context: To better understand and solve the problem, let's break down the given information and visualize the geometric relationships:

1. Triangles: There are two triangles involved in this problem:
   - Triangle ABC with vertices A, B, and C.
   - Triangle DEF with vertices D, E, and F.
2. Angles:
   - Angle A is given as 80 degrees.
   - Angle B is given as 60 degrees.
3. Parallel Lines:
   - Line DE is parallel to line BC.
4. Similar Triangles:
   - Since DE is parallel to BC, it can be inferred that triangle ABC is similar to triangle DEF. This means that the corresponding angles and sides are proportional.
5. Proportional Sides:
   - In similar triangles, the ratio of corresponding sides is constant. If we denote the sides of triangle ABC as a, b, and c (with a being opposite angle A, b opposite angle B, and c opposite angle C), and the sides of triangle DEF as d, e, and f (with d opposite angle D, e opposite angle E, and f opposite angle F), then the following proportions hold true:
   - \\(\\frac{a}{d} = \\frac{b}{e} = \\frac{c}{f}\\)
6. Angle CED:
   - We need to find the measure of angle CED. Since triangle ABC is similar to triangle DEF, the angles in these triangles are also proportional. Therefore, we can use the given angles A and B to determine the corresponding angles in triangle DEF.

By understanding these relationships and visualizing the geometric setup, you can use the properties of similar triangles and the given angle measures to determine the size of angle CED.

Hence, the simplified question becomes as follows:
In the figure, it is known that angle A = 80.0, angle B = 60.0, DE parallel BC, then the size of angle CED is:

Choices:

- A: 40°
- B: 60°
- C: 120°
- D: 140°

Please first conduct reasoning, and then answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.

Since DE is parallel to BC, triangle ADE is similar to triangle ABC by the Basic Proportionality Theorem (also known as Thales' Theorem). This means that the corresponding angles of these triangles are equal.

Given:

- Angle A = 80°
- Angle B = 60°

In triangle ABC:

- Angle C = 180° - Angle A - Angle B
  = 180° - 80° - 60°
  = 40°

Since DE is parallel to BC, Angle ADE is equal to Angle ABC, and Angle AED is equal to Angle ACB.

Therefore:

- Angle ADE = Angle ABC = 60°
- Angle AED = Angle ACB = 40°

In triangle ADE:

- Angle DAE = Angle A = 80°
- Angle ADE = 60°
- Angle AED = 40°

Now, Angle CED is an exterior angle for triangle ADE at vertex E. The exterior angle is equal to the sum of the two non-adjacent interior angles.

Thus:

- Angle CED = Angle ADE + Angle DAE
  = 60° + 80°
  = 140°

The correct option is:

D: 140°

Therefore, the correct answer is D.

Figure 16: An instance of input (highlighted in magenta) and Output for fine-tuning objective of `SpatialMath`.

A geometry problem often includes three components besides the answer choices: descriptive information, an implicit property, and an essential condition.

Descriptive Information refers to the directly observable and clearly portrayed content in the diagram, such as the presence of geometric shapes or intersection points of functions.

The second part, implicit property, involves a higher level of visual perception to discern from the diagram. It signifies strong visual conditions for problem-solving, such as the parallelism and perpendicularity between lines, the similarity and congruence among triangles, and the category and periodicity of functions.

The third part, essential condition, denotes the specific numerical or algebraic measurements.

Here is an incomplete textual question along with an image representing an associated diagram of a geometry problem where the implicit property is given in the text. Can you complete and write this question accurately in text-only form, taking into account the descriptive information and essential condition presented through the diagram? \nIncomplete Textual Question: As shown in the figure, DE parallel BC, then the size of angle CED is ()

Choices:

- A: 40°
- B: 60°
- C: 120°
- D: 140°

Associated Diagram: <image>

Complete Textual Question

Figure 17: Inference prompt for SpatialMath-SX$_{text_only}$ based fine-tuned SC model along with an example problem.

As shown in the figure, it is known that angle A = 80.0, angle B = 60.0, then the size of angle CED is ()
Choices:

- A: 40°
- B: 60°
- C: 120°
- D: 140°

<image>

Please generate the additional context for above question which may help in better comprehension of the problem. Context may include elaboration and simplification of the problem such that it becomes easier to solve it. Finally, reconstruct the question in simplified text-only form.

Figure 18: Inference prompt for SpatialMath-SX based fine-tuned SC model along with an example problem.

To better understand and solve the problem, let's break down the given information and relationships between the angles and sides in the given diagram:

1.Triangle ABC:
This is a triangle with vertices A, B, and C. The angle at vertex C is given as 80°.

2.Line DE parallel to BC:
This means that line DE is parallel to line BC. When two lines are parallel, they have the same slope or gradient, which implies that the triangles formed by these lines will have similar properties.

3.Angle BED:
This angle is given as 60°. Since DE is parallel to BC, we can use the properties of parallel lines to determine other angles in the diagram.

4.Triangles ADE and BEC:
These are the smaller triangles formed by the intersection of line DE with triangle ABC. The angle at vertex E is shared by both triangles, so it is common to both ADE and BEC.

5.Angles in Triangle ADE:
Since DE is parallel to BC, the angles at D and E in triangle ADE are equal (as they are corresponding angles). Let's call this common angle x.

6.Angles in Triangle BEC:
Similarly, since DE is parallel to BC, the angles at B and E in triangle BEC are equal. Let's call this common angle y.

7.Total Angle Sum in Triangle ABC:
The sum of the angles in any triangle must be 180°. Therefore, we can set up an equation using the given angles and the unknowns x and y.

By understanding these relationships and properties, you can use the given angles and the angle sum property of triangles to find the size of angle CED.

Simplified Question:

In the figure, DE is parallel to BC, then the size of angle CED is:

Choices:

- A: 40°
- B: 60°
- C: 120°
- D: 140°

Please first conduct reasoning and then answer the question and provide the correct option letter (e.g., A, B, C, D) at the end.

Figure 19: Inference prompt for default and `SpatialMath-RX` based fine-tuned reasoning solver model considering PI from downstream SC model along with an example problem.

Hyperparameter	Value
Batch Size	16
Gradient Accumulation	16
Learning Rate	4.0e-5
Weight Decay	0.01
Optimizer	AdamW
Scheduler	Linear
Warmup Steps / Ratio	50 steps
Max Sequence Length	32768
Mixed Precision	bf16
Inference: Model Parameters	
temperature	1.0
top_p	1.0
top_k	50

Table 11: Phi4 Hyperparameters.

Hyperparameter	Value
Batch Size	1
Gradient Accumulation	1
Learning Rate	1e-5
Weight Decay	0.0
Optimizer	AdamW
Scheduler	Cosine
Warmup Ratio	0.03
Max Sequence Length	32768
Mixed Precision	bf16
Inference:Model Parameters	
temperature	0.7
top_p	0.8
top_k	20
repetition_penalty	1.05

Table 12: LlaVa Next 34B Hyperparameters.