MathFlow: Enhancing the Perceptual Flow of MLLMs for Visual Mathematical Problems

Anonymous ACL submission

Abstract

Despite strong results on many tasks, multimodal large language models (MLLMs) still underperform on visual mathematical problem solving, especially in reliably perceiving and interpreting diagrams. Inspired by human problem-solving, we hypothesize that the ability to extract meaningful information from diagrams is pivotal, as it directly conditions subsequent inference. Hence, we introduce FlowVerse, a comprehensive benchmark that provides a fine-grained evaluation of MLLMs' perception and reasoning capabilities. Our preliminary results on FlowVerse reveal that existing MLLMs exhibit substantial limitations when extracting essential information and reasoned properties from diagrams and performing complex reasoning based on these visual inputs. In response, we introduce MathFlow, a modular problem-solving pipeline that decouples perception and inference into distinct stages, thereby optimizing each independently. Given the perceptual limitations observed in current MLLMs, we trained MathFlow-P-7B as a dedicated perception model. Experimental results indicate that MathFlow-P-7B yields substantial performance gains when integrated with various closed-source and open-source inference models. This demonstrates the effectiveness of the MathFlow pipeline and its compatibility with diverse inference frameworks. Our benchmark and code will be made publicly available.

1 Introduction

By enabling seamless interaction between visual data and natural language, multimodal large language models (MLLMs) (Radford et al., 2021; Li et al., 2024a; Chen et al., 2023a) are excelling in various tasks, including captioning (Lin et al., 2023), visual question answering (Bai et al., 2023) and visual dialogue (Liu et al., 2024a). Despite their impressive performance across diverse tasks, MLLMs have yet to fully demonstrate their potential in visual mathematical problem-solving (Chen et al., 2022; Lu et al., 2023), particularly in accurately perceiving and interpreting diagrams within these problems (Zhang et al., 2024c).

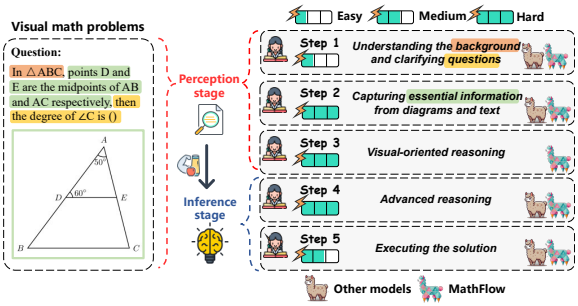


Figure 1: **The Typical Process of Humans Solving Visual Mathematical Problems.** We can summarize two key capabilities observed in the typical human problem-solving process: perception and inference. The perception capability involves extracting relevant information from both visual and textual inputs, ensuring accurate reasoning, which inspired the development of FlowVerse and MathFlow.

The typical process of humans for solving visual mathematical problems consists of five sequential steps (Krawec, 2014), as illustrated in Fig. 1. We categorize them into two key stages: *1) the perception stage* and *2) the inference stage*. The perception stage focuses on extracting relevant information from both visual and textual inputs that can be seamlessly integrated with the original problem statement (Polya and Pólya, 2014). In contrast, the inference stage concentrates on mathematical reasoning. Regarding the perception stage, existing MLLMs often struggle with mathematical images (Rahmanzadehgervi et al., 2024; Chen et al., 2021) and extract unreliable information from visual contents, which inevitably hampers the subsequent inference stage. Regarding the inference stage, MLLMs often struggle to generate coherent solutions when essential perceptual information is limited or unreliable during complex problem solving (Zhou et al., 2024; Qiao et al., 2024). We propose that prior methods' limited capabil-

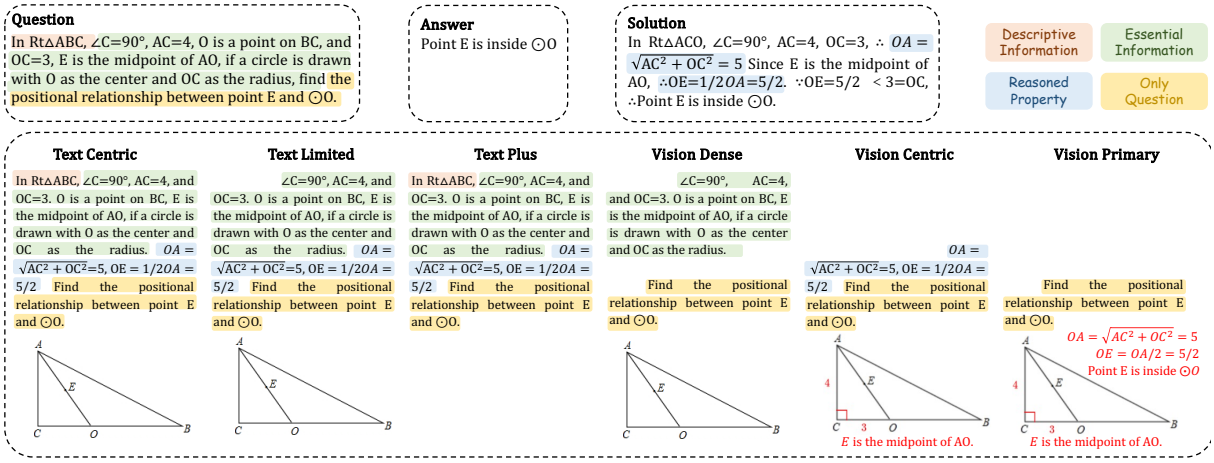


Figure 2: **Six Versions of Problems in FlowVerse.** FlowVerse begins by categorizing the original problem information into four distinct components: Descriptive Information (DI), Essential Information (EI), Only Question (OQ), and Reasoned Property (RP). The first three components are derived directly from the original problem statement, while RP is extracted from the solution and represents the inferences needed to solve the problem. In the Vision Centric version, we convert the EI into diagrams, while in the Vision Primary version, we convert both the EI and RP into diagrams.

ity to extract information during the perception stage bounds the overall problem-solving performance. To validate this assumption, we developed a new benchmark named FlowVerse. FlowVerse categorizes all the information used during problem-solving into four parts (Schoenfeld, 1987): *Descriptive Information (DI)*, which provides fundamental details describing the composition of the problem; *Essential Information (EI)*, which involves inferring critical details from both diagrams and text; *Only Question (OQ)*, representing the specific question posed in the text; and *Reasoned Property (RP)*, which refers to essential properties derived from diagrams and text during the problem-solving process rather than explicitly being provided in the problem. We design six problem variants using different combinations of information types, as illustrated in Fig. 2.

Our preliminary results on FlowVerse reveal that existing MLLMs struggle significantly in extracting EI and RP from diagrams, as well as in carrying out complex reasoning on given images. Motivated by these observations, we introduce a modular problem-solving pipeline, MathFlow, which explicitly decouples the problem-solving pipeline into two distinct stages: a perception stage and an inference stage. Fig. 4 illustrates that EI is first extracted from diagrams, followed by preliminary reasoning to derive RP in the perception stage. They are then converted into a text-level representation, further concatenated with the original problem's textual information to create a unified, enriched input. The inference model can easily generate a reasonable solution using the combined visual-derived and textual information.

Preliminary experiments on FlowVerse highlight the critical importance of robust perception capabilities in deriving useful properties from visual mathematical diagrams. However, a model as advanced as GPT-4V may exhibit deficiencies in this specific aspect (Rahmanzadehgervi et al., 2024; Feng et al., 2024; Ahn et al., 2024; Wang et al., 2024a). To address this, we propose MathFlow-P-7B, a specialized model designed to extract EI and RP information from visual diagrams, thereby facilitating visual mathematical problem-solving. The training of MathFlow-P-7B involves two stages: the multi-task pretraining stage and the supervised fine-tuning stage. The multi-task pretraining stage includes two core tasks—captioning EI and RP. The supervised fine-tuning stage further refines the model's perception capabilities for enhanced performance. Experimental results indicate that MathFlow-P-7B enables superior performance improvements when integrated with different inference models, underscoring the effectiveness of the MathFlow pipeline. Our contributions are summarized as follows: *1) FlowVerse*: We introduce a comprehensive benchmark specifically designed to evaluate the visual mathematical problem-solving capabilities of MLLMs across its meticulously crafted six problem versions. *2) MathFlow*: We propose a modular problem-solving pipeline that decouples the problem-solving process into perception and

inference, not only enhancing the model's ability to extract and reason with multimodal information effectively, but also empowering LLMs to handle visual mathematical problems. *3) MathFlow-P-7B*: We develop a perceptive mathematics model, an MLLM specifically optimized for visual mathematical problems, demonstrating SOTA performance across various benchmarks.

2 Related Work

Visual Mathematical Reasoning with MLLMs. Solving visual mathematical problems (*e.g.*, geometry diagrams, algebraic plots, and *etc*) requires both strong reasoning ability and accurate interpretation of visual primitives and symbolic content (Yan et al., 2024; Qiao et al., 2025; Zhang et al., 2025). Most previous works are dedicated to improving the reasoning process, including chain-of-thought strategies (Xu et al., 2025; Deng et al., 2024), tool-aided reasoning (Trinh et al., 2024; Chen et al., 2023b), test time scaling (Wang et al., 2025b; Hosseini et al., 2024), and reinforcement learning (Wang et al., 2025a; Jiang et al., 2025). Several recent works (Guo et al., 2025; Jia et al., 2024; Wei et al., 2025) suggest that one of the major bottlenecks in visual mathematical reasoning is inaccurate visual comprehension.

Mathematics Evaluation Benchmarks. The use of LLMs and MLLMs for solving visual mathematical problems has been extensively explored in several studies (Elhenawy et al., 2024; Wang et al., 2024c; Yang et al., 2024; Peng et al., 2024). GPTs (Floridi and Chiriatti, 2020; OpenAI, 2023b,a). To better assess the progress of LLMs and MLLMs in mathematical solving and to drive further improvement, a number of mathematics-specific evaluation benchmarks have been introduced (Liu et al., 2024b; Satpute et al., 2024; Zhang et al., 2024a). However, many mathematical problems include diagrams, which has led researchers to increasingly focus on developing MLLMs that can effectively handle both textual and visual content (Gao et al., 2023a; Li et al., 2024b). To provide a more comprehensive and in-depth evaluation of MLLMs, MathVerse (Zhang et al., 2024c) builds upon MathVista, focusing on eliminating textual redundancy to ensure genuine interpretation of visual diagrams rather than reliance on textual shortcuts. It also introduces a Chain-of-Thought (CoT) evaluation strategy (Wei et al., 2022; Wu et al., 2024) for a detailed assessment of intermediate reasoning steps. However, our analysis reveals that these benchmark datasets still have limitations.

3 FlowVerse Benchmark

To validate MLLMs' capabilities regarding perception and inference, we first propose FlowVerse, which categorizes problem information into several key components.

Dataset Composition. FlowVerse comprises 2,000 visual mathematical problems collected from real exam questions in both Chinese and English, resulting in over 12,000 test samples. *Detailed statistics for the dataset composition are presented in Tab. 7 in the Appendix.* To guarantee the dataset's quality and precision, we conducted a thorough review to verify the accuracy of the answers and analyses, as well as to ensure consistency between the questions and their corresponding diagrams. Notably, to avoid any confusion, we named the initially collected dataset **FlowVerse**†, while the subsequent dataset containing six versions of the problem is called **FlowVerse**.

Furthermore, to achieve a comprehensive evaluation, this meticulously curated dataset encompasses four key components of information flow: Descriptive Information (DI), Essential Information (EI), Reasoned Property (RP), and Only Question (OQ). Each of these components is carefully crafted to provide distinct insights into the factors that influence the perception capabilities of MLLMs: ❶ ***Descriptive Information*** (DI) provides fundamental details about the composition of figures, referring directly to the observable and explicitly depicted elements within a diagram, such as geometric shapes or intersection points in functions. These descriptions help establish context and frame the problem. ❷ ***Essential Information*** (EI) refers to the critical details required for problem-solving, such as specific values or relationships between geometric elements (e.g., $\angle A = 45°$, $AD \perp BC$). FlowVerse incorporates information directly from visual diagrams as part of EI, recognizing that in visual math problems, much of the essential information comes from the diagrams. Thus, accurately extracting EI from multiple modalities is crucial for MLLMs to solve problems effectively. ❸ ***Reasoned Property*** (RP) represents information inferred through higher-level visual abstraction and reasoning combined with relevant mathematical knowledge. Unlike EI, RP requires understanding beyond simple visual perception, as diagrams may not explicitly

Table 1: **Mathematical Evaluation on Six Problem Versions in FlowVerse**. DI, EI, RP, OQ refer to the textual or visual Descriptive Information, Reasoned Property, Essential Information, Only Question, respectively. The Text Plus Version does not involve image input. "CoT-E" or "Acc" denotes whether to employ the FlowVerse-CoT-E strategy or not. The highest accuracy for each group of MLLMs is marked in **bold**. *The full table is provided in the Appendix.*

Model	All		Text Centric DI+RP+EI+OQ		Text Limited RP+EI+OQ		Text Plus DI+RP+EI+OQ		Vision Dense EI+OQ		Vision Centric RP+EI+OQ		Vision Primary RP+EI+OQ	
	CoT-E	Acc	CoT-E	Acc	CoT-E	Acc	CoT-E	Acc	CoT-E	Acc	CoT-E	Acc	CoT-E	Acc
Open-source MLLMs														
InfiMM-Math-7B	36.5	28.8	43.8	38.1	40.6	36.7	46.1	40.1	28.8	15.4	39.6	30.3	26.1	23.2
InternVL2.5-8B	44.7	41.0	49.2	41.3	40.5	38.4	49.6	42.7	38.4	20.2	41.0	35.9	35.8	33.9
Qwen2.5-VL-7B	53.8	42.2	60.1	52.8	58.9	51.3	62.0	55.0	45.0	31.0	50.8	46.3	48.1	45.3
VLM-R1-7B†	50.7	41.2	59.0	54.2	57.9	49.8	65.5	58.9	36.2	24.5	46.1	37.8	30.6	26.1
Qwen2-VL-72B	52.3	48.6	59.4	47.3	54.3	45.7	63.7	50.0	40.8	25.3	50.9	42.1	47.6	37.0
InternVL2.5-78B	54.7	50.1	66.1	62.7	64.1	60.3	67.8	64.7	48.7	34.3	63.0	58.8	59.6	57.7
MathFlow$^{\star}_{\text{Qwen2.5-VL-7B}}$	57.0	46.0	62.0	53.8	60.8	52.2	64.2	56.0	49.0	39.1	54.5	51.6	52.0	51.5
Math-specialized MLLMs														
MathLLaVA-13B	38.0	29.9	45.1	39.3	44.4	37.4	–	–	36.2	18.6	41.7	35.9	37.0	34.2
MultiMath-7B	44.0	34.2	50.6	44.8	49.9	42.9	–	–	41.7	22.1	47.2	40.4	39.7	38.8
SVE-Math-Qwen2.5-7B	46.0	39.4	53.1	47.3	53.4	45.8	–	–	44.2	28.6	48.9	44.2	45.8	42.0
Closed-source MLLMs														
Qwen-VL-Max	43.0	36.3	49.8	42.1	46.7	38.3	53.9	51.0	38.6	15.2	42.7	33.2	29.6	27.8
GPT-4o-mini	51.3	44.5	58.7	54.8	58.2	53.2	59.6	55.2	41.1	26.0	57.4	50.1	49.7	47.6
Claude-3.5-Sonnet	56.9	49.6	60.8	52.6	58.7	50.3	64.0	58.3	45.0	25.4	56.5	48.0	48.1	45.2
GPT-4o	55.1	47.8	61.0	56.8	58.7	54.4	62.2	58.2	45.2	30.0	58.6	52.6	54.1	51.0
GPT-4V	56.2	53.4	69.1	57.1	65.0	55.0	72.0	61.4	48.1	30.3	61.8	46.3	42.0	36.7
Gemini-2.5-pro	62.0	55.3	68.3	61.9	66.1	60.8	68.9	64.1	52.1	37.1	65.7	57.9	57.0	54.6
GPT-5	65.8	60.1	74.3	68.1	73.5	66.7	77.0	69.2	53.8	44.7	67.1	61.7	60.3	57.5
MathFlow$^{\star}_{\text{GPT-5}}$	**66.5**	**61.8**	**74.6**	**68.5**	**73.8**	**67.2**	**77.0**	**69.3**	**58.2**	**54.1**	**70.2**	**66.7**	**69.4**	**66.2**

convey relationships between geometric elements. In FlowVerse, RP is often not directly present in the original problem but requires additional human annotation. These annotations serve as guiding tips to assist solvers, enhancing accuracy and enabling a more comprehensive and fine-grained evaluation of MLLMs' capabilities. ❹ ***Only Question*** (OQ) refers to the specific question posed in the text. Though a small part of the problem, it determines what needs to be solved, without which the solution process cannot proceed.

Based on the four categories, expert annotators systematically remove different types of information within questions and progressively incorporate critical elements into the diagrams. As illustrated in Fig 2, we generate six versions of each problem based on four key components of information, resulting in 12,000 test instances. With this curated problem set, we provide a comprehensive evaluation of the visual perception capabilities of MLLMs and assess whether these abilities can effectively support multimodal mathematical reasoning. The details of each problem version are: ❶ ***Text Centric Version*** retains all content, including DI, RP, EI, and OQ. Notably, in this version, we manually convert all RP and EI components into textual form to effectively analyze their impact on the final reasoning process across different modalities. ❷ ***Text Limited Version*** removes the DI from the Text Centric Version while retaining the other components. ❸ ***Text Plus Version*** excludes the image entirely

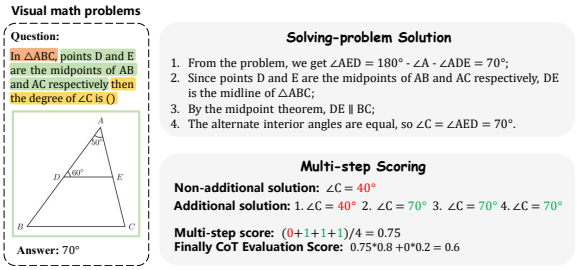


Figure 3: **The FlowVerse-CoT-E Strategy.**

from the Text Centric Version, reducing the input from a multimodal to a purely text-based format. ❹ ***Vision Dense Version*** removes the RP from the Text Limited Version. ❺ ***Vision Centric Version*** converts the EI from text to image, starting with the Text Limited Version, thereby making it more visually focused. ❻ ***Vision Primary Version*** further converts the Reasoned Property from text to image, based on the Vision Centric Version, resulting in an entirely visual input for analysis.

FlowVerse-CoT-E: Evaluation Strategy. As previously mentioned, several studies have evaluated the abilities of MLLMs from a Chain-of-Thought (CoT) perspective (Wei et al., 2022). However, these approaches typically rely on GPT or human intervention to decompose the model's response into sequential steps and subsequently evaluate the correctness of each step. The variability of intermediate steps complicates this process, as GPT's interpretation and decomposition of questions can introduce errors (termed MathVerse-CoT-E), leading to significant noise and undermining the accuracy and

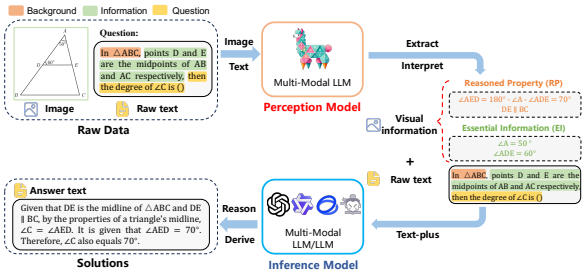


Figure 4: **The Overview of MathFlow Pipeline.** To effectively train MLLMs for problem-solving, we decouple MLLMs into two sub-modules: the perception model and the inference model. The perception model aims to extract and interpret visual information, converting it into a form that can be effectively processed. The inference model uses this extracted information, with the original question, to reason and derive solutions.

consistency of subsequent evaluations (Bai et al., 2023; Dong et al., 2024). To address these challenges, we propose a **robust CoT-based evaluation strategy**, termed FlowVerse-CoT-E, which sequentially integrates components of an authoritative problem-solving solution into the prompts. Specifically, experts first develop an authoritative problem-solving method for each question in FlowVerse and then decompose it into several solution steps. These steps are progressively incorporated into the prompts provided to the MLLMs for reasoning. As illustrated in Fig. 3, the results generated by MLLMs under different prompts are aggregated using a weighted approach to derive the final outcome. Notably, to ensure the robustness of the evaluation, we provide multiple solutions for questions that admit more than one valid approach. The scoring process is formulated as follows:

$$\text{Score}_{final} = \alpha\left(\frac{1}{N}\sum_{i=1}^{N}\text{Score}_i\right) + (1-\alpha)\text{Score}_0, \tag{1}$$

where α defines as a balancing factor between intermediate reasoning steps, setting $\alpha = 0.8$ by default to highlight the importance of CoT reasoning. Score_i represents the MLLM's score up to step i, while Score_0 reflects its score based solely on the final answer, without intermediate steps.

Key Findings on FlowVerse. As reported in Tab. 1, our preliminary experiments on FlowVerse benchmark highlight the critical importance of robust perception capabilities in deriving useful properties from visual mathematical diagrams. ***Please refer to §5.2 for more detailed discussions.***

4 MathFlow: A Modular Visual Mathematical Problem-Solving Pipeline

Motivated by the key findings illustrated above, we aim to enhance the extraction of accurate textual EI and RP. Hence, we introduce a modular problem-solving pipeline, MathFlow. The MathFlow pipeline consists of two stages to alleviate the limitations of MLLMs in visual mathematical problem-solving. In the perception stage, the model extracts critical information from visual data and converts it into a text representation, integrating it with the original problem. This enriched input is then passed to the inference stage.

A model's perception and inference capabilities are inherently distinct, with strong perception being a prerequisite for accurate inference. Consequently, we prioritize enhancing the model's perception abilities to enable precise extraction and interpretation of visual information, thereby supporting more reliable inference. Then, the inference stage can utilize other state-of-the-art MLLMs/LLMs, allowing the best inference capabilities to complement the perception improvements brought by MathFlow. This modular methodology ensures flexibility and robustness in solving complex visual mathematical problems. For simplicity, we refer to the model in the perception stage as the "perception model" and the model in the inference stage as the "inference model".

Furthermore, MathFlow's training strategy is divided into two stages: the multi-task pretraining stage and the supervised fine-tuning stage. Upon completion of training, we obtain a fine-tuned MLLM (MathFlow-P-7B).

Multi-Task Pretraining Stage. The training tasks in this stage primarily include the EI caption task and the visual-oriented reasoning task.

On the one hand, the EI caption task aims to train the perception model to generate textual descriptions of essential elements directly from visual inputs. For this, we fine-tune a pretrained Qwen2-VL-7B model (Wang et al., 2024b) by utilizing datasets like MAVIS (Zhang et al., 2024d) and Geo170k (Gao et al., 2023b), which contain real or synthetic visual mathematics image-text pairs.

On the other hand, the visual-oriented reasoning task is designed to extract higher-level abstractions and relationships, requiring the model to engage in abstract reasoning—such as deducing relationships between geometric shapes or identifying intersection points in a function. To support this task,

we developed a training dataset called ***MathFlow-RP***, sampled from educational materials that include detailed problem-solving solutions. ***Details of MathFlow-RP are presented in the Appendix.*** Specifically, we deconstructed each problem's corresponding solution into multiple steps, using the previous step as context to guide the MLLM in predicting the next step of the solution.

Finally, we performed mixed training on both tasks, ensuring that the data ratio between the two tasks was maintained at 3:1. During this stage, the LLM backbone is frozen, and training is focused on the perceiver resampler and vision encoder module to ensure that the visual components can effectively extract and process the necessary information.

Supervised Fine-Tuning Stage. In this stage, our goal is to enhance the model's response quality and adapt it more effectively to the current task context. To achieve this, we meticulously developed a supervised fine-tuning dataset called ***MathFlow-SFT***, focusing on retaining only specific numerical values, relationships between geometric elements, and essential information necessary for defining the solution space. *Details of this process are provided in the Appendix.* During this stage, we freeze the vision encoder and concentrate on training the perceiver resampler and LLM backbone. This approach aims to improve the model's ability to accurately interpret visual inputs and generate concise textual representations crucial for problem-solving.

5 Experiments

5.1 Experimental Setup

Datasets. We used both publicly available datasets and internal data to support training. For the EI caption task, we began by filtering out images with excessively high resolutions and those unrelated to mathematical content. Ultimately, we obtained 650,000 image-text pairs. For the visual-oriented reasoning task, we first selected 40,000 problems from our custom question bank that contained detailed solution processes. We then decomposed each problem's solution into individual steps and incorporated them into prompts, resulting in a final dataset of 130,000 samples. *Please refer to the Appendix for more details.*

Baselines. We benchmarked MathFlow against SOTA MLLMs. For *open-source* models, we compared MathFlow against Qwen-VL-Plus (Bai et al., 2023), Qwen2-VL (Wang et al., 2024b), InfiMM-Math (Han et al., 2024), InternVL-2.5 (Chen et al., 2024) and Qwen-VL-Max (Bai et al., 2023). For *close-source* models, we included Claude-sonnet-3.5 (Anthropic, 2024), Gemini1.5-Pro (Team et al., 2024),SPHINX-MoE (Gao et al., 2024), InternLM-XC2 (Dong et al., 2024) and GPT-4V (OpenAI, 2023b). For *math-specialized model*, we included SVE-Math (Zhang et al., 2025), MultiMath (Peng et al., 2024), MathLlava (Shi et al., 2024).

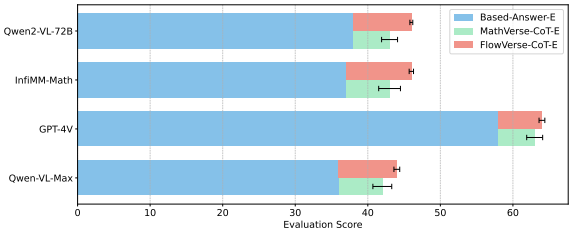


Figure 5: **Robustness Comparison of Two Different CoT Evaluation Strategies on FlowVerse†.**

Implementation Details. All experiments were conducted in a zero-shot setting to demonstrate the generalization capabilities of MLLMs in mathematical reasoning without relying on few-shot prompting or any additional fine-tuning. By default, we employ th CoT-based) prompting technique (Wei et al., 2022), which encourages MLLMs to perform complete reasoning steps for a fine-grained evaluation. We conduct all experiments on NVIDIA A100 GPUs. During the training of MathFlow, we consistently used DeepSpeed Zero2. In the multi-task pretraining stage, we adopted a maximum learning rate of 1e-5. In the supervised fine-tuning stage, we used a maximum learning rate of 5e-6. *Please refer to the Appendix for more details.*

5.2 Benchmark Results on FlowVerse

Tab. 1 compares several state-of-the-art (SOTA) MLLMs on FlowVerse benchmark, including both closed-source and open-source models. Notably, in FlowVerse—where Reasoned Property (RP) and Essential Information (EI) information are manually annotated—MathFlow does not require additional extraction of these components. Therefore, we do not include MathFlow in direct comparisons for FlowVerse. We mainly analyze the performance by the FlowVerse-CoT-E evaluation and derive the following observations:

MLLMs Rely More on *Reading text* than *Seeing Diagrams*. When comparing the Text Centric and Text Plus versions in FlowVerse, we observe that most MLLMs improve in performance when visual input is removed, such as a +4.3% increase for Qwen2-VL-72B. Conversely, when comparing

Table 2: **Mathematical Evaluation on Six Problem Versions in MathVerse's *testmini* Set.** The "All" score is calculated without including the average of the Text Only version. "CoT-E" or "Acc" indicates the use of the proposed CoT evaluation strategy or not. ⋆ denotes that MathFlow-P-7B functions as the perception model. The highest accuracy for closed-source and open-source MLLMs is marked in **blod** and uline respectively.

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
Qwen-VL-Plus (Bai et al., 2023)	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 (Team et al., 2023)	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 (Bai et al., 2023)	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 (OpenAI, 2023b)	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
MathFlow$^{\star}{}_{\mathrm{GPT-4V}}$	**56.7**	**43.8**	**65.2**	**51.1**	**58.9**	**46.4**	**62.1**	**48.5**	**53.7**	**40.3**	**52.1**	**37.4**	**52.5**	**39.0**
SPHINX-MoE (Gao et al., 2024)	25.8	15.6	33.3	22.2	21.9	16.4	40.7	18.3	21.1	14.8	19.6	12.6	18.3	9.1
InternLM-XC2 (Dong et al., 2024)	25.9	16.5	36.9	22.3	28.3	17.0	42.5	16.5	20.1	15.7	24.4	16.4	19.8	11.0
MAVIS-7B (Zhang et al., 2024d)	27.5	-	41.4	-	29.1	-	42.5	-	27.4	-	24.9	-	14.6	-
InfiMM-Math (Han et al., 2024)	34.5	-	46.7	-	32.4	-	-	-	38.1	-	32.4	-	15.8	-
Qwen2-VL-72B (Wang et al., 2024b)	38.9	28.5	49.2	35.4	35.6	27.8	52.1	39.6	38.7	26.4	35.1	28.5	22.5	13.2
MathFlow$^{\star}{}_{\mathrm{Qwen2-VL-72B}}$	40.5	31.7	52.3	39.3	37.7	31.7	55.2	45.5	40.5	30.3	37.2	32.4	24.6	17.1

the Text Centric and Text Limited versions, we notice a significant drop in performance, such as a -4.1% decrease for GPT-4V. These results emphasize the necessity of both comprehensive textual descriptions and effective visual encoding for robust problem-solving in MLLMs. *Please refer to the Appendix for more results.*

MLLMs Significantly Benefit from *EI*. Incorporating the EI within diagrams challenges MLLMs to accurately identify and interpret these critical components visually during mathematical problem-solving. The Vision Centric results show a significant drop in performance for most MLLMs compared to their accuracy in the Text Limited version. In other words, we conclude that a truly robust MLLM must effectively extract EI from diagrams and visual data. This ability to independently comprehend and interpret visual elements is crucial for successfully solving visual mathematical problems, indicating that enhanced perception capabilities are a key factor in improving MLLMs' performance.

MLLMs Significantly Benefit from *RP*. By removing the RP from the question text, we observe a significant decline in accuracy (Text Limited Version *vs* Vision Dense Version) for most MLLMs, particularly in open-source models. This indicates that RP contains many inferred relationships that are essential for guiding the reasoning process and enhancing the overall problem-solving capabilities of MLLMs. Furthermore, when we convert RP from the text modality to the image modality, the performance of most MLLMs decreases substantially. This not only reinforces the observation that MLLMs rely more on *reading text* than *interpreting diagrams* but also highlights that MLLMs significantly benefit from the inclusion of *RP*.

The FlowVerse-CoT-E Strategy Shows More Robust. In Fig. 5, we compare two different

Table 3: **Performance Comparison of MLLMs on MathVerse and FlowVerse† Datasets. FlowVerse†** indicates the raw version of the dataset. "$MathFlow^{*}$" denotes the MathFlow pipeline, whose perception model is MathFlow-P-7B. *The full table is provided in the Appendix.*

Model	**MathVerse**	**FlowVerse†**
Qwen2-VL-72B	38.9	52.3
InternVL-2.5-78B	43.2	54.7
GPT-4V	54.4	56.2
Claude-sonnet-3.5	57.4	64.0
Gemini 2.5-pro	59.9	68.9
MathFlow$^{\star}{}_{\mathrm{Qwen2-VL-72B}}$	48.1	58.3
MathFlow$^{\star}{}_{\mathrm{GPT-4V}}$	56.7	59.3
MathFlow$^{\star}{}_{\mathrm{InternVL-2.5-78B}}$	56.8	60.1
MathFlow$^{\star}{}_{\mathrm{Claude-sonnet-3.5}}$	60.8	68.1
MathFlow$^{\star}{}_{\mathrm{Gemini2.5-pro}}$	**62.4**	**70.4**

CoT evaluation strategies: MathVerse-CoT-E and FlowVerse-CoT-E. Specifically, we evaluated them using both MathVerse-CoT-E and FlowVerse-CoT-E methods on FlowVerse† dataset, repeating 5 times. The results show that while MathVerse-CoT-E can demonstrate the effectiveness of fine-grained assessment, its evaluation lacks robustness due to its reliance on GPT to perform key-step extraction on MLLM responses, followed by multi-step scoring. In contrast, FlowVerse-CoT-E uses manually deconstructed steps, resulting in more accurate and robust evaluations.

5.3 MathFlow Analysis

Overall Performance. Tab. 1 and Tab. 2 show the performance of MLLMs on FlowVerse and MathVerse's testmini set. MathFlow achieves the highest overall accuracy outperforming other models like GPT-4V. In terms of CoT-based evaluation (CoT-E), MathFlow$^{\star}{}_{\mathrm{GPT-5}}$ also demonstrates consistent superiority. On the other hand, Tab. 3 shows the performance comparison of MLLMs on the FlowVerse† and MathVerse datasets, where

Table 4: **Ablation Analysis of MathFlow on FlowVerse†. Bold** numbers indicate the best performance. *The full table is provided in the Appendix.*

Perception Model for EI	Perception Model for RP	Inference Model	COT-E (%)
Qwen2-VL-2B	GPT-4V	GPT-4V	49.0
Qwen2-VL-7B	GPT-4V	GPT-4V	53.3
InternVL2.5-8B	GPT-4V	GPT-4V	53.8
Qwen2.5-VL-7B	GPT-4V	GPT-4V	53.9
MathFlow-P-7B	GPT-4V	GPT-4V	**58.9**
GPT-4V	Qwen2-VL-2B	GPT-4V	49.8
GPT-4V	Qwen2-VL-7B	GPT-4V	52.9
GPT-4V	MathFlow-P-7B	GPT-4V	**57.3**
GPT-4V	GPT-4V	GPT-4V	56.2
MathFlow-P-7B	MathFlow-P-7B	GPT-4V	**59.3**
MathFlow-P-7B	MathFlow-P-7B	GPT-4o	69.6
MathFlow-P-7B	MathFlow-P-7B	Gemini 2.5-pro	70.4
MathFlow-P-7B	MathFlow-P-7B	DeepSeek-v3	73.4
MathFlow-P-7B	MathFlow-P-7B	DeepSeek-r1	**75.6**

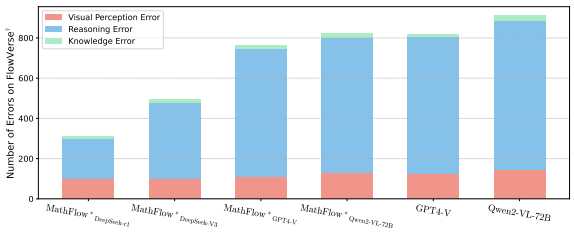


Figure 6: **Comparison of Error Distributions Across Models on FlowVerse†.**

FlowVerse† refers to the raw, unmodified version of FlowVerse. Notably, MathFlow$^{\star}{}_{\mathrm{GPT-5}}$ achieves the highest accuracy across both datasets. These results not only highlight MathFlow's adaptability but also underscore the critical role of perception capabilities. Furthermore, we also evaluate MathFlow$^{\star}{}_{\mathrm{Qwen2.5-VL-7B}}$ against a set of math-specialized MLLMs on Tab. 1, MathFlow-P-7B consistently outperforms these baselines across all six problem versions, demonstrating its superior adaptability to diverse multimodal inputs. We attribute these gains to its explicit decoupling of the perception and inference stages, which particularly ensures accurate perception.

Error Analysis. Fig. 6 categorizes errors types into three types. Experts annotated the responses of multiple models on FlowVerse† to identify these error types. We observe that MathFlow* not only shows an overall reduction in error rate compared to the corresponding base model but also exhibits a decrease in visual perception errors. Moreover, even with MathFlow$^{\star}{}_{DeepSeek-r1}$, a significant number of perception errors still occur, underscoring our assertion that perception capability is crucial.

Table 5: **Performance comparison of various models trained with MathFlow as the perception model.**

Model	mPLUG-Doc	Qwen	InternVL	mPLUG-Doc*	MathFlow-P-7B
GPT-4o	47.8	58.7	59.1	62.1	**69.6**
DS-R1	59.3	67.8	68.6	70.2	**75.6**

Ablation Studies. Tab. 4 presents the ablation analysis of MathFlow on the FlowVerse† dataset, examining the contributions of different perception models for EI and RP, as well as the impact of varying inference models. The analysis demonstrates that when GPT-4V is used as the inference model, MathFlow consistently outperforms other configurations . Furthermore, we found that substituting other SOTA LLMs as inference models yielded even higher performance, with MathFlow$^{\star}{}_{DeepSeek-r1}$ reaching its peak performance.

We further compare more models trained with MathFlow as the perception model in Tab. 5, and the MathFlow-trained variant mPLUG-Doc*—in which the EI and RP perception modules share a common backbone but differ in inference model In all cases, these models underperformed MathFlow-P-7B, underscoring the efficacy and robustness of the MathFlow pipeline.

This remarkable effectiveness of MathFlow stems from its explicit separation of perception and inference, allowing for independent optimization of the perception model and resulting in more accurate visual feature extraction. By converting complex visual information into textual representations that are readily consumable by inference models, MathFlow significantly enhances the reasoning capabilities of various LLMs/MLLMs. Notably, we emphasize that MathFlow enables SOTA LLMs to **solve visual mathematical problems without any additional training**, effectively extending their capabilities beyond pure language processing to visual mathematical reasoning. *Please refer to the Appendix for more analysis.*

6 Conclusion

We introduce FlowVerse to investigate the bottlenecks of MLLMs in visual mathematical problem-solving, which reveals the limitations in current MLLMs' ability to accurately perceive and process visual information. Motivated by these findings, we propose MathFlow, which decouples the problem-solving pipeline into perception and inference stages. Given the poor perception performance, we developed MathFlow-P-7B using a two-stage training strategy.

Limitations

In FlowVerse, we initially categorized the collected mathematical problems based on subjects and subfields that reflect varying degrees of multimodal content. These classification methods enable a comprehensive evaluation of MLLM capabilities across different dimensions. However, an additional categorization by difficulty level—similar to the approach used in datasets like MATH (Hendrycks et al., 2021) and WeMath (Qiao et al., 2024), where problems are sorted by complexity—would provide deeper insights into model performance. This extra layer of differentiation could significantly enhance the model's assessment, and we plan to explore this in our future work.

Furthermore, the issues with FlowVerse primarily stem from the fact that they are mainly available in English and Chinese. This limitation restricts the model's applicability to a wider range of languages. By expanding the dataset to include a greater diversity of languages, we could enhance the multilingual capabilities of MLLMs and create benchmarks that are more inclusive for users who speak languages other than English and Chinese.

References

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.

Anthropic. 2024. claude-3-5-sonnet system card.

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

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746*.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P Xing, and Liang Lin. 2021. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *arXiv preprint arXiv:2105.14517*.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023a. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023b. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Linger Deng, Yuliang Liu, Bohan Li, Dongliang Luo, Liang Wu, Chengquan Zhang, Pengyuan Lyu, Ziyang Zhang, Gang Zhang, Errui Ding, Yingying Zhu, and Xiang Bai. 2024. R-cot: Reverse chain-of-thought problem generation for geometric reasoning in large multimodal models. *arXiv preprint arXiv:2410.17885*.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, and 1 others. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Mohammed Elhenawy, Ahmad Abutahoun, Taqwa I Alhadidi, Ahmed Jaber, Huthaifa I Ashqar, Shadi Jaradat, Ahmed Abdelhay, Sebastien Glaser, and Andry Rakotonirainy. 2024. Visual reasoning and multi-agent approach in multimodal large language models (mllms): Solving tsp and mtsp combinatorial challenges. *arXiv preprint arXiv:2407.00092*.

Tony Haoran Feng, Paul Denny, Burkhard Wuensche, Andrew Luxton-Reilly, and Steffan Hooper. 2024. More than meets the ai: Evaluating the performance of gpt-4 on computer graphics assessment questions. In *Proceedings of the 26th Australasian Computing Education Conference*, pages 182–191.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and 1 others. 2023a. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and 1 others. 2023b. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*.

Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, and 1 others. 2024. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*.

Zixian Guo, Ming Liu, Qilong Wang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. 2025. Integrating visual interpretation and linguistic reasoning for math problem solving. In *ICCV*.

Xiaotian Han, Yiren Jian, Xuefeng Hu, Haogeng Liu, Yiqi Wang, Qihang Fan, Yuang Ai, Huaibo Huang, Ran He, Zhenheng Yang, and 1 others. 2024. Infimm-webmath-40b: Advancing multimodal pre-training for enhanced mathematical reasoning. *arXiv preprint arXiv:2409.12568*.

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. *arXiv preprint arXiv:2103.03874*.

Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. 2024. V-star: Training verifiers for self-taught reasoners. *arXiv preprint arXiv:2402.06457*.

Mengzhao Jia, Zhihan Zhang, Wenhao Yu, Fangkai Jiao, and Meng Jiang. 2024. Describe-then-reason: Improving multimodal mathematical reasoning through visual comprehension training. *arXiv preprint arXiv:2404.14604*.

Chaoya Jiang, Yongrui Heng, Wei Ye, Han Yang, Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. 2025. Vlm-r^3: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. *arXiv preprint arXiv:2505.16192*.

Jennifer L Krawec. 2014. Problem representation and mathematical problem solving of students of varying math ability. *Journal of Learning Disabilities*, 47(2):103–115.

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

Zhihao Li, Yao Du, Yang Liu, Yan Zhang, Yufang Liu, Mengdi Zhang, and Xunliang Cai. 2024b. Eagle: Elevating geometric reasoning through llm-empowered visual instruction tuning. *arXiv preprint arXiv:2408.11397*.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, and 1 others. 2023. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yan Liu, Renren Jin, Ling Shi, Zheng Yao, and Deyi Xiong. 2024b. Finemath: A fine-grained mathematical evaluation benchmark for chinese large language models. *arXiv preprint arXiv:2403.07747*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

OpenAI. 2023a. Chatgpt. https://chat.openai.com.

OpenAI. 2023b. GPT-4V(ision) system card.

OpenAI. 2023c. Introducing openai o1.

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

George Polya and George Pólya. 2014. *How to solve it: A new aspect of mathematical method*, volume 34. Princeton university press.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, and 1 others. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*.

Runqi Qiao, Qiuna Tan, Peiqing Yang, Yanzi Wang, Xiaowan Wang, Enhui Wan, Sitong Zhou, Guanting Dong, Yuchen Zeng, Yida Xu, Jie Wang, Chong Sun, Chen Li, and Honggang Zhang. 2025. We-math 2.0: A versatile mathbook system for incentivizing visual mathematical reasoning. *Preprint*, arXiv:2508.10433.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. 2024. Vision language models are blind. *arXiv preprint arXiv:2407.06581*.

Ankit Satpute, Noah Gießing, André Greiner-Petter, Moritz Schubotz, Olaf Teschke, Akiko Aizawa, and Bela Gipp. 2024. Can llms master math? investigating large language models on math stack exchange. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2316–2320.

Alan H Schoenfeld. 1987. Pólya, problem solving, and education. *Mathematics magazine*, 60(5):283–291.

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

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Qwen Team. 2024. Qwen2.5-llm: Extending the boundary of llms.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482.

Karen D Wang, Eric Burkholder, Carl Wieman, Shima Salehi, and Nick Haber. 2024a. Examining the potential and pitfalls of chatgpt in science and engineering problem-solving. In *Frontiers in Education*, volume 8, page 1330486. Frontiers Media SA.

Peiyu Wang, Yichen Wei, Yi Peng, Xiaokun Wang, Weijie Qiu, Wei Shen, Tianyidan Xie, Jiangbo Pei, Jianhao Zhang, Yunzhuo Hao, and 1 others. 2025a. Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning. *arXiv preprint arXiv:2504.16656*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yikun Wang, Siyin Wang, Qinyuan Cheng, Zhaoye Fei, Liang Ding, Qipeng Guo, Dacheng Tao, and Xipeng Qiu. 2025b. Visuothink: Empowering lvlm reasoning with multimodal tree search. *arXiv preprint arXiv:2504.09130*.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024c. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*.

Haoran Wei, Youyang Yin, Yumeng Li, Jia Wang, Liang Zhao, Jianjian Sun, Zheng Ge, Xiangyu Zhang, and Daxin Jiang. 2025. Slow perception: Let's perceive geometric figures step-by-step. *Preprint*, arXiv:2412.20631.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhaolong Wu, Abul Hasan, Jinge Wu, Yunsoo Kim, Jason PY Cheung, Teng Zhang, and Honghan Wu. 2024. Chain-of-though (cot) prompting strategies for medical error detection and correction. *arXiv preprint arXiv:2406.09103*.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. Llava-cot: Let vision language models reason step-by-step. In *ICCV*.

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. *arXiv preprint arXiv:2412.11936*.

Zhen Yang, Jinhao Chen, Zhengxiao Du, Wenmeng Yu, Weihan Wang, Wenyi Hong, Zhihuan Jiang, Bin Xu, Yuxiao Dong, and Jie Tang. 2024. Mathglm-vision: Solving mathematical problems with multi-modal large language model. *arXiv preprint arXiv:2409.13729*.

Boning Zhang, Chengxi Li, and Kai Fan. 2024a. Mario eval: Evaluate your math llm with your math llm–a mathematical dataset evaluation toolkit. *arXiv preprint arXiv:2404.13925*.

Di Zhang, Jianbo Wu, Jingdi Lei, Tong Che, Jiatong Li, Tong Xie, Xiaoshui Huang, Shufei Zhang, Marco Pavone, Yuqiang Li, and 1 others. 2024b. Llama-berry: Pairwise optimization for o1-like olympiad-level mathematical reasoning. *arXiv preprint arXiv:2410.02884*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and 1 others. 2024c. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, and 1 others. 2024d. Mavis: Mathematical visual instruction tuning. *arXiv preprint arXiv:2407.08739*.

Shan Zhang, Aotian Chen, Yanpeng Sun, Jindong Gu, Yi-Yu Zheng, Piotr Koniusz, Kai Zou, Anton van den Hengel, and Yuan Xue. 2025. Open eyes, then reason: Fine-grained visual mathematical understanding in mllms. *Preprint*, arXiv:2501.06430.

Zihao Zhou, Shudong Liu, Maizhen Ning, Wei Liu, Jindong Wang, Derek F Wong, Xiaowei Huang, Qiufeng Wang, and Kaizhu Huang. 2024. Is your model really a good math reasoner? evaluating mathematical reasoning with checklist. *arXiv preprint arXiv:2407.08733*.

A Table of Contents

This appendix presents detailed descriptions of FlowVerse and MathFlow, including dataset construction, training strategies, extended experiments, qualitative examples, and discussions of limitations and future work. A structured breakdown is shown below.

- **Sec. A: Additional Details of FlowVerse**
  - B.1 Data Curation
  - A.2 Subject and Subfield Definition
  - A.3 Detailed Statistics of FlowVerse
  - A.4 Evaluation on FlowVerse
  - A.5 Qualitative Examples
- **Sec. B: Additional Details of MathFlow**
  - B.1 Details of MathFlow-RP
  - B.2 Details of MathFlow-SFT
  - B.3 More Analysis of Experiment Results
  - B.4 Structured Caption Evaluation
  - B.5 More Detailed Experiment Results
  - B.6 Image-Only Scenario
  - B.7 Reasoning Efficiency Across Model Scales
  - B.8 Qualitative Examples
- **Sec. C: Limitation and Future Work**

B Additional Details of FlowVerse

B.1 Data Curation

Data Collection. We comprehensively collect visual math problems from internal data[1,2]. Specifically, we focused on collecting problems related to plane geometry, functions, and algebra. To further enhance the diversity of the dataset, we introduced additional subcategories within each main category, capturing more nuanced types of problems. After the initial collection, we compiled approximately 2,000 visual math problems.

Data Categorization and Review. Initially, human annotators categorized the collected problems into three main subjects: plane geometry, solid geometry, and functions. Within each subject, the problems were further subdivided into twelve fine-grained categories. Subsequently, we meticulously reviewed the dataset, manually correcting problems with incorrect answers and discarding those that contained multiple diagrams, visual solutions, or had content overly similar to other problems. After this careful curation, we preserved 2,000 high-quality math problems with paired diagrams for FlowVerse, covering a wide range of subjects and subfields. Notably, to ensure the quality and richness of the dataset, we excluded problems with minimal Descriptive Information (DI) during the dataset creation process. Notably, to facilitate evaluation, all proof-based questions have been reformulated into a question-and-answer format, ensuring consistency and comparability across different problem types.

Table 6: **Statistics of FlowVerse.**

Statistic	**Number**
Total Questions	2000
- Subjects/subfields	3/15
- Multiple-choice questions	848 (62.4%)
- Free-form questions	1,152 (37.6%)
- **Questions with solutions**	**2,000 (100%)**
Multiple-choice question	
- Proportion of answer A	171 (20.2%)
- Proportion of answer B	257 (30.3%)
- Proportion of answer C	198 (23.3%)
- Proportion of answer D	200 (23.6%)
- Proportion of answer E&F	22 (2.6%)
Number of unique images	1,906 (95.3%)
Number of unique questions	2,000 (100%)
Number of unique answers	561 (28.1%)
Number of English questions	400 (20%)
Number of Chinese questions	1600 (80%)
Maximum question length	769
Maximum answer length	351
Average question length	104.1
Average answer length	9.9

Transformation of Problem Versions. Given the four distinct types of information embedded within the questions, human annotators meticulously transform each problem into six different versions, as detailed in Sec.3.1 of the main paper. For this purpose, we utilize NetPad[3] to annotate diagrams for the Text Dominant, Vision Centric, and Vision Primary versions.

For the Text Centric version, as depicted in Fig. 7, where the raw data presents both EI and RP fully within the diagrams, we selectively remove portions of this information from the diagrams and integrate them into the text. This approach emphasizes the textual representation of the key elements.

In the Vision Centric version, NetPad is employed to translate EI from text into a visual representation, while the corresponding textual information is removed from the Text Centric version.

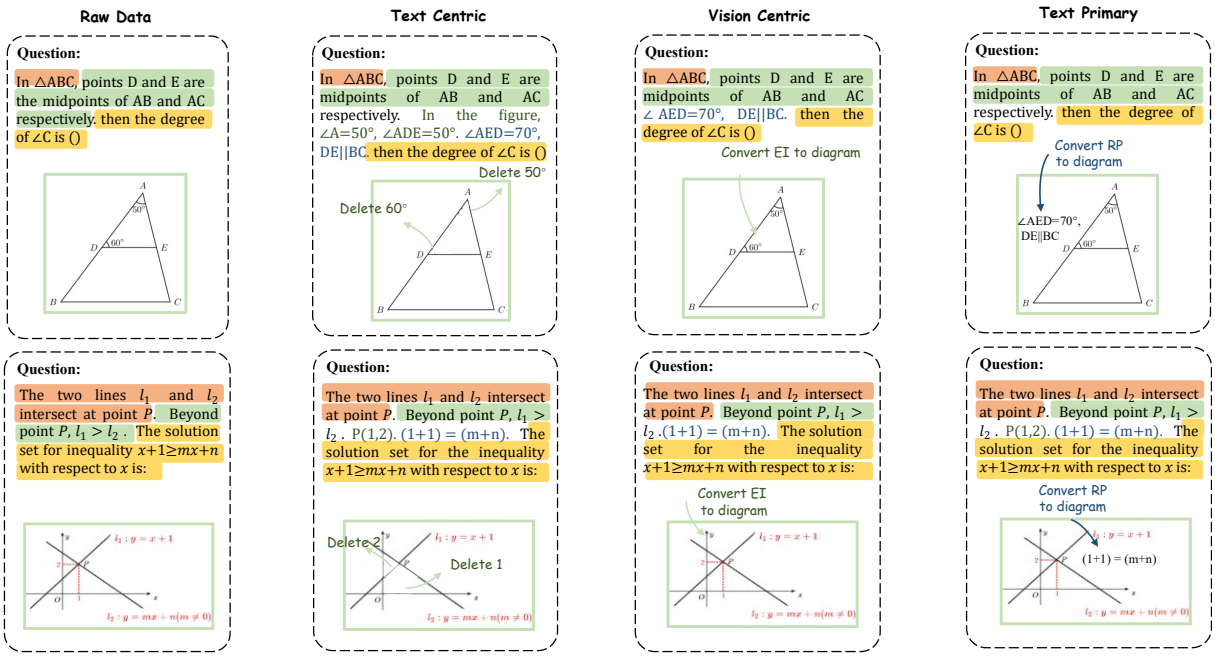


Figure 7: **Manual Modification for EI in FlowVerse.** For the original problems shown, we transfer some of the EI from diagrams to question texts (highlighted in green) to mark the Vision Centric version.

Table 7: **Length of Different Problem Versions in FlowVerse.**

Problem Version	Character
Text Centric & Text Plus	
- Maximum question length	962
- Maximum answer length	351
- Average question length	153.3
- Average answer length	9.9
Text Limited	
- Maximum question length	937
- Maximum answer length	351
- Average question length	123.9
- Average answer length	9.9
Vision Dense	
- Maximum question length	688
- Maximum answer length	351
- Average question length	74.7
- Average answer length	9.9
Vision Centric	
- Maximum question length	571
- Maximum answer length	351
- Average question length	75.8
- Average answer length	9.9

This transformation aims to shift the balance of information representation towards visual content.

For the Vision Primary version, NetPad is further utilized to convert RP from text into visual form based on the Vision Centric version, ultimately creating a version in which all pertinent information is conveyed solely through diagrams.

B.2 Subject and Subfield Definition

Plane Geometry. This foundational area studies the properties and relationships of points, lines, and surfaces within a two-dimensional plane. It covers key concepts such as circles, triangles, parallelograms, and parallel lines, providing a comprehensive context to evaluate the spatial reasoning and logical deduction abilities of MLLMs. Furthermore, expert annotators have categorized the problems into twelve fine-grained categories, as illustrated in Fig. 8 highlighting various dimensions of visual mathematical skills.

- **Circles.** This subfield involves understanding properties and relationships associated with circles, such as radius, diameter, tangents, arcs, chords, and inscribed angles. Evaluating MLLMs in this area tests their ability to comprehend and reason about the unique attributes and properties that define circular geometry.

- **Triangles.** This category examines various properties of triangles, including side lengths, angles, congruence, similarity, and the Pythagorean theorem. Assessing models of learning and language (MLLMs) on triangles helps evaluate their understanding of fundamental geometric principles and their ability to solve problems involving different types of triangles, such as equilateral, isosceles, and

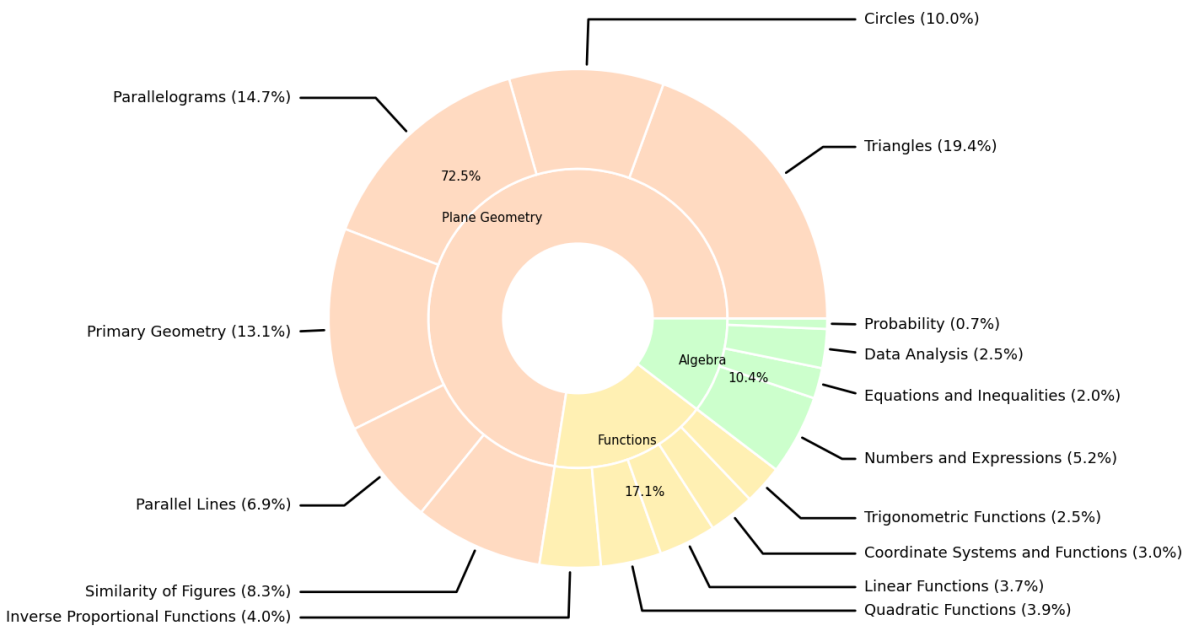


Figure 8: **Subject Distribution of FlowVerse.**

scalene triangles.

- **Parallelograms.** This subfield focuses on understanding the properties of parallelograms. It covers the relationships between opposite sides and angles, the concept that the diagonals bisect each other, and specific types of parallelograms such as rectangles, rhombuses, and squares. It also tests the understanding of the characteristics that define and differentiate these shapes.

- **Parallel Lines.** This subfield focuses on the study of parallel lines, especially the angles created when they are intersected by a transversal. Key angles to examine include alternate interior angles and corresponding angles. It is essential to evaluate the ability of MLLMs to apply these properties in order to solve complex geometric problems.

- **Similarity of Figures.** This area examines the criteria for similarity between geometric figures, such as side ratios and corresponding angles. It requires MLLMs to apply proportional reasoning and recognize conditions under which two figures are similar, testing their understanding of geometric relationships in a broader context.

- **Primary Geometry.** This subfield deals with fundamental concepts such as points, lines, and basic shapes. It lays the foundation for understanding more complex geometrical relationships, testing MLLMs' ability to grasp and reason with the basic building blocks of geometry.

Functions. This encompasses analyzing mathematical relationships between variables, ranging from simple evaluations of function values to more complex tasks like examining different function types and their behaviors. We assess MLLMs' capabilities using four categories of function problems, as exemplified in Fig 9 .

- **Inverse Proportional Functions:** This category involves functions where the product of two variables remains constant. Evaluating MLLMs on inverse proportional functions helps assess their understanding of non-linear relationships and their ability to apply these principles in problem-solving.

- **Quadratic Functions:** Problems in this category include analyzing the properties of quadratic equations, such as identifying the vertex, roots, axis of symmetry, and the impact of changing coefficients. MLLMs are tested on their ability to understand and manipulate quadratic expressions to solve for function behavior.

- **Linear Functions:** Linear functions represent relationships with a constant rate of change.

This category focuses on understanding slope and y-intercept and interpreting graphical representations of linear equations, which are essential for evaluating MLLMs' grasp of fundamental linear relationships.

- **Coordinate Systems and Functions:** This subfield deals with understanding how functions are represented within coordinate planes, including plotting points, interpreting graphs, and understanding shifts and transformations of basic functions. It assesses MLLMs' skills in translating functional relationships into visual formats.

- **Trigonometric Functions:** Problems in this category involve understanding trigonometric relationships, such as sine, cosine, and tangent, and their applications to solve problems involving angles and periodic phenomena. Evaluating MLLMs on these functions helps gauge their understanding of trigonometric concepts and their application in different scenarios.

Algebra. encompasses a wide range of topics that focus on understanding mathematical relationships through symbols and expressions. It involves solving problems related to probability, data analysis, equations, and numerical expressions. We evaluate MLLMs' skills using four distinct categories within algebra:

- **Probability:** This category covers fundamental probability concepts, including calculating the likelihood of events, understanding random experiments, and applying basic probability laws. It assesses MLLMs' ability to handle uncertainty and predict outcomes based on given conditions.

- **Data Analysis:** Problems in this category involve interpreting and analyzing data sets, including calculating measures such as mean, median, mode, and range, and understanding data representations like charts and graphs. Evaluating MLLMs in this area tests their ability to draw meaningful conclusions from quantitative data.

- **Equations and Inequalities:** This subfield includes solving linear and non-linear equations, as well as inequalities involving one or more variables. MLLMs are evaluated on their ability to manipulate and solve algebraic expressions and apply them to real-world scenarios.

- **Numbers and Expressions:** This category deals with understanding numerical properties, simplifying algebraic expressions, and performing arithmetic operations. It tests MLLMs' grasp of numerical relationships and their ability to work with algebraic symbols effectively.

B.3 Detailed Statistics of FlowVerse

More Data Statistics. In Tab. 6, we present a detailed breakdown of the FlowVerse dataset. It is important to note that all problems in FlowVerse are sourced from internet data rather than other datasets, and each question is annotated manually. Additionally, our evaluation aims to effectively showcase the reasoning capabilities of MLLMs with moderate-level mathematical knowledge without restricting their performance through overly complex domain-specific theorems or extensive commonsense knowledge. Consequently, we focus on problems at the high school level, deliberately excluding advanced college-level topics such as calculus and graph theory.

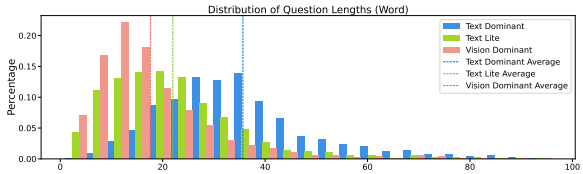


Figure 9: **Distribution of Question Length for Four Problem Versions.** We present the distribution of question length for the four problem versions, with the horizontal axis representing question length in characters and the vertical axis depicting the corresponding probability distribution.

Problem Length Variance. In Tab. 7, we present the variations in question and answer lengths across five versions of problems in FlowVerse, excluding the Vision Primary category, as it contains only the "Only Question" component. By removing predefined components such as Descriptive Information (DI), Reasoned Property (RP), and Essential Information (EI), the maximum and average lengths of questions decrease progressively, whereas the answer lengths remain unaffected. Fig. 9 visualizes the character-level variation in question length for four problem versions: Text Centric (blue), Text Limited (green), Vision Centric (red), and Vision Dense (yellow). As DI, EI, and RP are sequentially

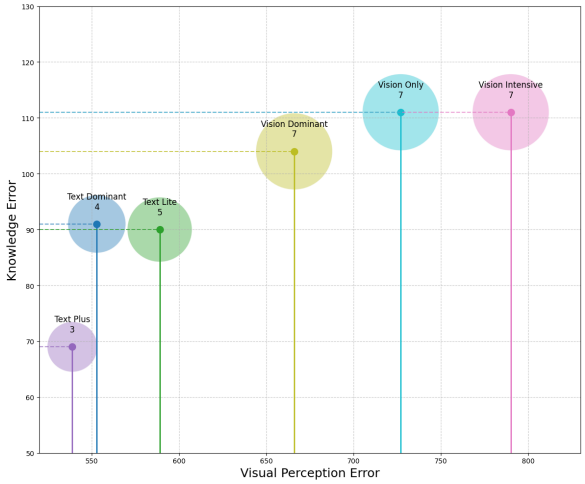


Figure 10: **Visualization of Different Error Type across Different Versions using GPT-4 on FlowVerse.** The horizontal axis represents different problem versions, while the vertical axis indicates the error types. The radius of each bubble corresponds to the number of visual perception errors, with smaller radii indicating fewer visual perception errors.

omitted from the Text Centric version, we observe a clear downward trend in both the distribution of questions lengths and their average values.

B.4 Details of COT Evaluation on FlowVerse

Prompt Design for Response Generation. We employ two distinct types of prompts for free-form and multiple-choice questions, respectively, as per the guidelines established by MathVerse (Zhang et al., 2024c), which are detailed in Tab. 8. To effectively elicit the Chain-of-Thought (CoT) reasoning capabilities of MLLMs, we also incorporate the phrase *"first conduct reasoning"* to encourage a more structured and logical approach to problem-solving.

Prompt for CoT Evaluation. As discussed earlier, our proposed Chain-of-Thought evaluation, termed FlowVerse-CoT-E, is specifically designed to assess the reasoning depth and perceptual accuracy of MLLMs in visual mathematics problems. FlowVerse-CoT-E utilizes a structured prompt that guides the model step-by-step through the problem-solving process, ensuring a more transparent evaluation of the model's ability to logically infer and connect various aspects of the problem. This evaluation mechanism distinguishes between different reasoning paths taken by the model, thereby providing a nuanced understanding of its intermediate problem-solving capabilities, as shown in Tab. 9.

Error Analysis of GPT-4 on FlowVerse. Fig. 10 illustrates GPT-4V's Knowledge Error, Reasoning Error, and Visual Perception Error across different versions of problem representations. The Text Plus version demonstrates the lowest Knowledge and Reasoning Errors, along with a smaller Visual Perception Error, as indicated by its relatively small bubble radius. This makes it the most effective approach for minimizing overall errors. In contrast, the Vision Centric, Vision Only, and Vision Dense versions show significantly higher Reasoning and Knowledge Errors, accompanied by the largest bubble radii, indicating substantial Visual Perception Error. This suggests that an over-reliance on visual information increases the complexity of both perception and reasoning. The Text Centric and Text Limited versions fall in between, exhibiting moderate error rates and bubble sizes, which implies a slight increase in perception difficulty as the representation shifts away from textual information. Overall, the analysis highlights that enriched textual representation, such as in the Text Plus version, is crucial for achieving optimal performance in minimizing Knowledge, Reasoning, and Visual Perception errors, underscoring the importance of a balanced information representation for effective problem-solving. In conclusion, the result further confirms that the limited capabilities of prior methods to extract information during the perception stage restrict the overall problem-solving performance, which also guides us in the development of the modular problem-solving pipeline, MathFlow.

B.5 Qualitative Examples

Figs. 11-13 illustrate the differences across six versions of problem representations: Text Centric, Text Limited, Text Plus, Vision Centric, Vision Intensive, and Vision Primary. Each version provides a distinct balance between textual and visual information, thereby influencing the accessibility and interpretability of the problem.

The Text Centric version includes detailed descriptive information along with key reasoning steps in the text, while the Text Limited version reduces some descriptive elements to focus on core content. The Text Plus version enriches this further with added visual context for better comprehension. On the other hand, the Vision Centric and Vision Dense versions emphasize visual elements by gradually minimizing textual information, whereas the Vision Primary version conveys almost all the information through diagrams with minimal textual support.

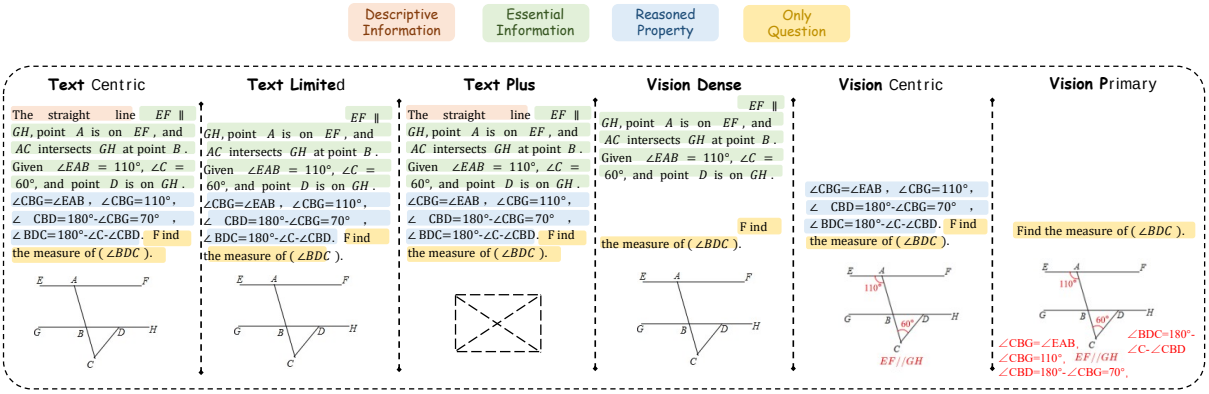


Figure 11: **Comparison of Six Problem Versions in FlowVerse.**

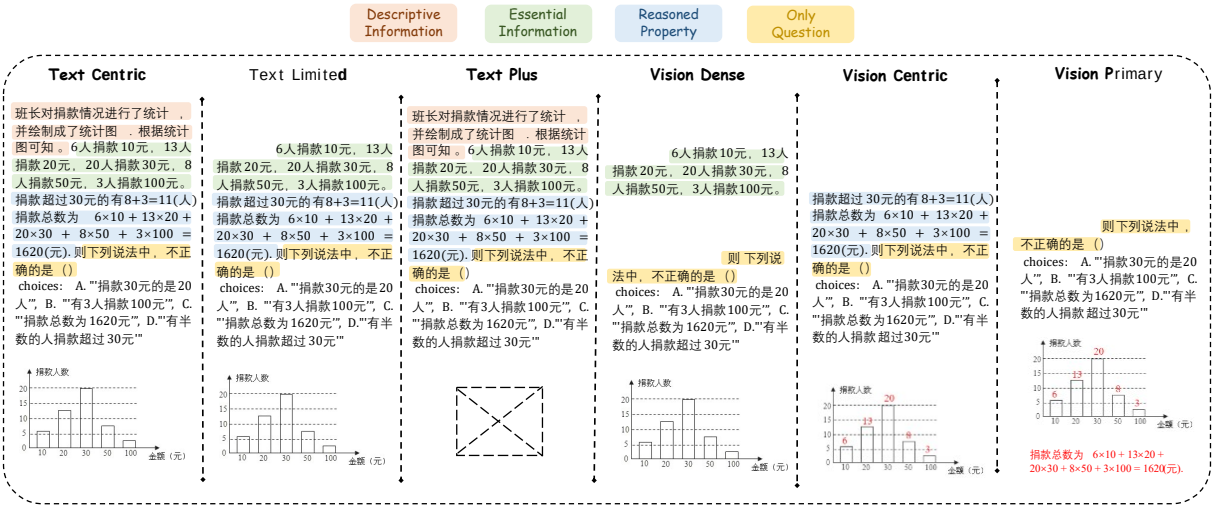


Figure 12: **Comparison of Six Problem Versions in FlowVerse.**

Table 8: **Input Prompt of MLLMs for Response Generation.**

Question	**Prompt**
Free-form Question	Please first conduct reasoning, and then answer the question and provide the final value, e.g., 1, 2.5, 300, at the end. – **Question:** $\{question\}$
Multiple-choice Question	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:** $\{question\}$

These examples highlight the impact of information representation on the model's ability to comprehend and solve problems, demonstrating how variations in text and visual content can influence both perception and reasoning capabilities.

C Additional Details of MathFlow

C.1 Details of MathFlow-RP

As mentioned in Sec.3.2, We designed MathFlow-RP to enhance the reasoned property of MathFlow during the multi-task pretraining stage for the RP caption task, which focuses on extracting higher-level abstractions and relationships, requiring the model to engage in abstract reasoning—such as

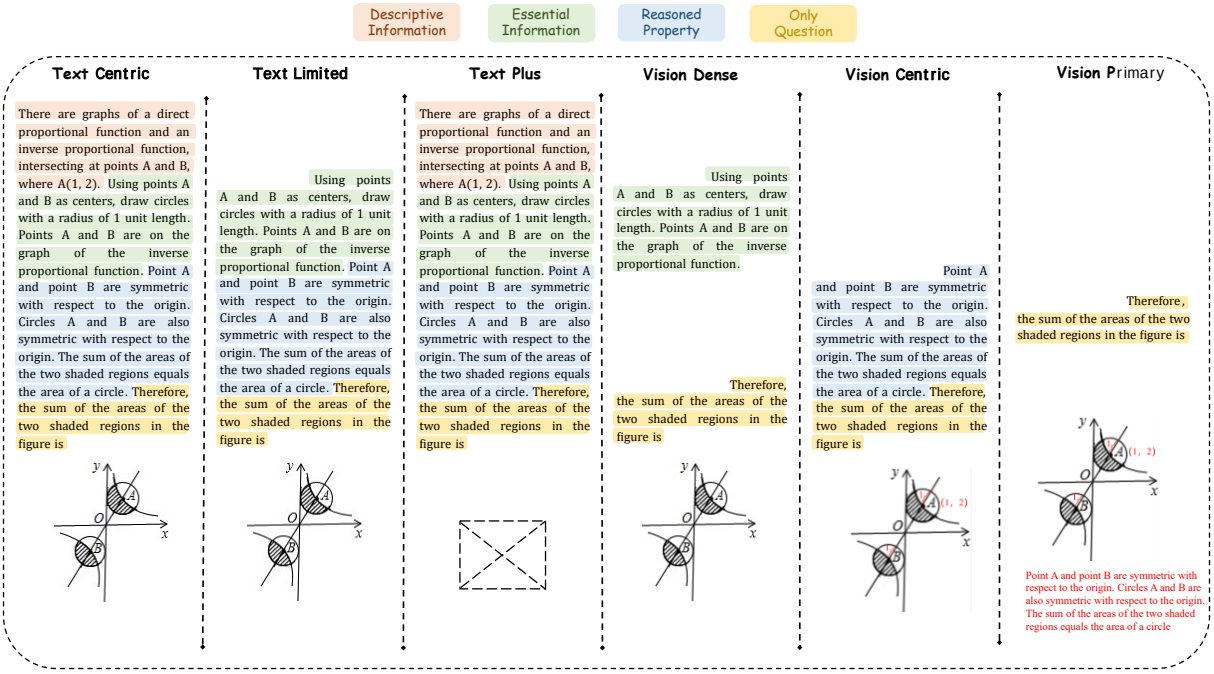


Figure 13: **Comparison of Six Problem Versions in FlowVerse.**

Table 9: **Configuration for the FlowVerse-CoT-E.**

Input	**Prompt**
Question & Sequence of Solutions	You will be provided with a visual mathematics problem along with a detailed solution procedure. Your task is to solve the problem by utilizing the provided solution steps as guidance. Here are examples: – Question: XXX – Solution: 1. XXX 2. XXX 3. XXX – Final Solution: XXX Here is what you need to solve: – **Question:** $\{question\}$ – **Solution:** $\{solution\}$ – **Final Solution:**

deducing relationships between geometric shapes or identifying intersection points in a function.

Data Collection. As previously described, the MathFlow-RP dataset is meticulously curated to enhance the model's response quality and adapt it more effectively to the current task context. This dataset consists of 40,000 questions, each accompanied by a detailed solution, thus providing a robust foundation for enhancing the model's capability to comprehend complex visual mathematical reasoning.

Data Annotation and Filtering. Specifically, as illustrated in Fig. 15, Qwen2.5-72B (Team, 2024) is employed to extract the key steps from these solutions, given the prohibitive cost of utilizing GPT-4 for the extraction across such a large dataset. As a result, we were able to compile a total of 13,000 annotated instances, which serve as the core training data for enhancing the model's problem-solving capabilities. Notably, since some problems involve multiple questions, we divide them into different instances for clarity.

Training Strategy. Building on this approach, the extracted solution steps are sequentially appended to the original question, enabling the model to predict the subsequent solution step based on the cumulative information. This iterative process facilitates a chain-of-thought reasoning mechanism by

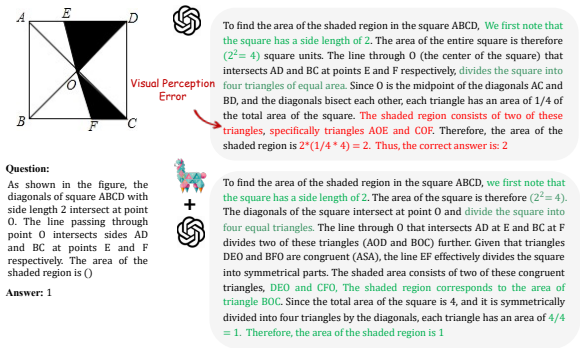


Figure 14: **Problem-solving Comparison of MathFlow$^{\star}_{\text{GPT-4V}}$ and GPT-4V.** The results indicate that MathFlow-P-7B demonstrates significantly better capability in understanding and addressing complex mathematical problems, showcasing more effective perception and inference abilities compared to GPT-4V.

progressively integrating previous steps into the prompt, allowing the model to incrementally build upon earlier conclusions. Specifically, each step in the solution is derived using pertinent properties, such as congruent triangles or properties of parallel lines, ensuring that every next prediction is grounded in logical progression. By extracting, labeling, and integrating key steps, the model is trained to emulate human-like deductive reasoning, ultimately enhancing its problem-solving accuracy for complex visual mathematical tasks.

C.2 Details of MathFlow-SFT

As mentioned in Sec.3.2, the MathFlow-SFT (Supervised Fine-Tuning) dataset is meticulously curated to enhance the model's response quality and adapt it more effectively to the current task context during the supervised fine-tuning stage of MathFlow. The dataset focuses on extracting and refining the key visual and textual elements necessary for accurate mathematical reasoning, with specific emphasis on retaining only the essential information to clearly define solution steps.

Data Collection. The MathFlow-SFT dataset samples visual mathematical problems with detailed solutions, which are sourced from diverse educational resources, including textbooks, standardized exam questions, and synthetic problems specifically designed for visual challenges. Also, MathFlow-SFT is collected to ensure coverage across different types of mathematical reasoning, such as plane geometry, algebra, and trigonometry, to provide a comprehensive training base.

Annotation of Key Elements. To ensure that the model effectively perceives and interprets diagrams, as Fig. 16 shows, we manually annotated the critical components of each problem based on both the descriptive information provided in the problem statement and the corresponding solution. The annotations include:

- Geometric Elements: Labels for points, lines, angles, and other geometric properties.
- Relationships: Specific relationships, such as congruency or parallelism, are highlighted to help the model understand the underlying structure.
- Reasoned Property Extraction: Abstract properties inferred from the given diagrams and textual context are labeled, emphasizing the reasoning process needed to progress through each solution.

C.3 More Analysis of Experiment Results

Tab. 10 compares the performance of various closed and open-source Multimodal Large Language Models (MLLMs) on the MathVision dataset across different mathematical subjects. The highest scores are highlighted in red for closed-source and blue for open-source models.

MathFlow$^{\star}_{\text{GPT-4Turbo}}$ leads among closed-source models in arithmetic (Ari), logical reasoning (Log), and geometry (Angle), showing its strength in diverse mathematical domains. Among open-source models, MathFlow$^{\star}_{\text{Qwen2-VL-72B}}$ performs well in algebra (Alg), arithmetic (Ari), and topology (Topo), demonstrating robust capabilities in visual mathematical tasks.

Overall, MathFlow models perform competitively with both closed and open-source baselines, showcasing their adaptability to complex mathematical problem-solving by effectively integrating visual perception information and logical reasoning.

C.4 Structured Caption Evaluation

We evaluate the quality of visual captioning by assessing how accurately models extract essential elements and their relationships from mathematical diagrams. We adopt the *Structured F1* metric, which quantifies alignment between predicted and ground-truth structured outputs. For each diagram, we decompose the caption into two sets: **Essential Information (EI)**, such as geometric primitives (e.g., points, lines, circles, and labeled symbols),

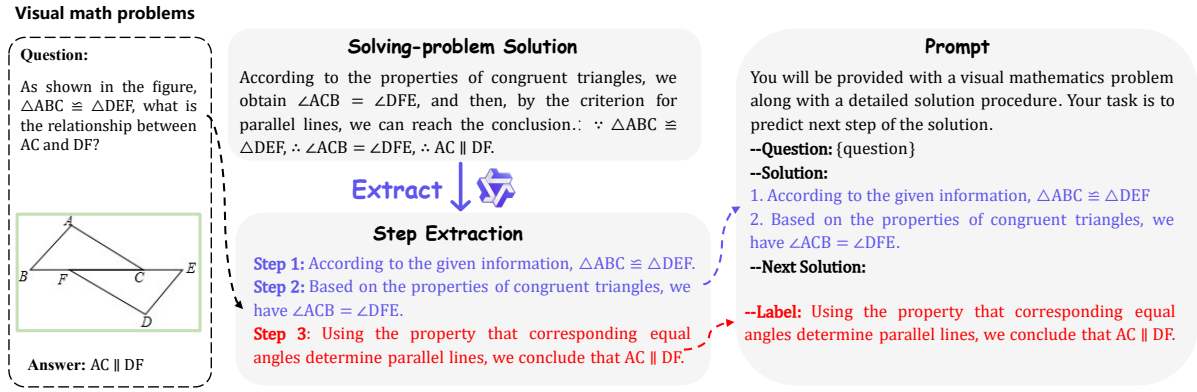


Figure 15: **Data Annotation of the MathFlow-RP** We first employ Qwen2.5-72B to extract the corresponding steps from the solving-problem solution, then select step *N* as the target for prediction. Subsequently, the preceding *N-1* steps are provided as input within the prompt, enabling the MLLM to predict the next step based on this sequential context.

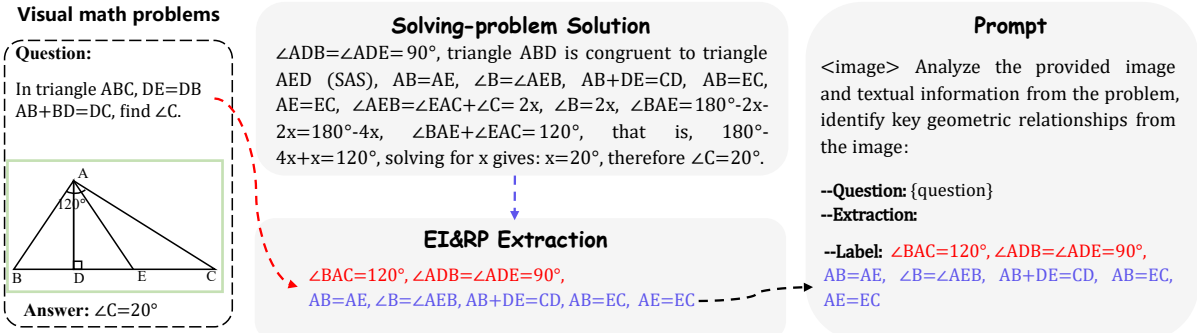


Figure 16: **Data Annotation of the MathFlow-SFT.**We manually extract the corresponding EI and RP from the solving-problem solution and associated diagram. In this representation, the red-highlighted portions indicate EI, while the blue-highlighted sections represent RP.

Table 10: **Comparison of model performances across various mathematical subjects on the MathVision dataset.** The highest accuracy for closed-source and open-source MLLMs is marked in red and blue respectively.

Model	Overall	Alg	AnaG	Ari	CombG	Comb	Cnt	DescG	GrphT	Log	Angle	Area	Len	SolG	Stat	Topo	TransG
Qwen-VL-Plus	10.72	11.3	17.9	14.3	12.7	4.8	10.5	15.4	8.9	14.3	11.6	6.4	10.0	14.3	6.9	8.7	11.31
Qwen-VL-Max	15.59	10.7	19.1	20.0	16.9	12.5	17.9	16.4	12.2	21.0	13.3	14.2	19.8	11.5	20.7	13.0	17.3
Gemini Pro	17.66	15.1	10.7	20.7	20.1	11.9	7.5	20.2	21.1	16.8	19.1	19.0	20.0	14.3	13.8	17.4	20.8
GPT-4Turbo	30.26	37.7	33.3	46.4	25.0	28.6	25.3	15.4	27.8	31.9	30.6	29.0	31.9	28.7	37.9	17.4	23.2
MathFlow$^{*}_{\text{GPT-4Turbo}}$	32.01	43.9	38.7	48.5	28.2	29.0	25.4	17.7	30.6	32.1	32.9	30.7	35.1	31.5	40.0	20.3	26.0
ShareGPT4V-13B	11.88	7.5	15.5	16.4	10.7	8.9	9.0	11.5	8.9	7.6	11.6	13.0	17.4	10.3	8.6	8.7	12.5
SPHINX-MoE	14.18	7.8	17.9	14.3	15.6	9.5	11.9	12.5	15.6	12.6	16.2	15.6	17.8	13.5	12.1	8.7	16.1
InternLM-XComposer2-VL	14.54	9.3	15.5	12.1	15.3	11.3	10.5	14.4	22.2	19.3	19.7	15.6	15.0	11.9	15.5	26.1	15.5
Qwen2-VL-72B	25.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
MathFlow$^{*}_{\text{Qwen2-VL-72B}}$	28.14	38.2	36.5	47.5	27.1	23.8	20.6	22.2	21.5	27.6	19.5	24.0	28.3	27.3	33.0	35.0	18.1

and **Reasoned Properties (RP)**, which represent higher-order relationships between elements (e.g., parallelism, intersection).

Each predicted EI or RP tuple is matched against human-annotated ground truth using exact set match. The precision, recall, and F1 are computed as:

$$\text{Precision} = \frac{|\text{Prediction} \cap \text{GroundTruth}|}{|\text{Prediction}|} \quad (2)$$

$$\text{Recall} = \frac{|\text{Prediction} \cap \text{GroundTruth}|}{|\text{GroundTruth}|} \quad (3)$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

This evaluation is conducted on a curated subset of FlowVerse where high-quality human annotations are available. As shown in Tab. 11, MathFlow-P-7B significantly outperforms general-purpose

Table 11: **Structured caption quality evaluation.** MathFlow-P-7B achieves superior performance in extracting both visual elements (EI) and their relationships (RP) from diagrams.

Model	EI (F1)	RP (F1)
GPT-4o-mini	84.4%	65.3%
GPT-4o	87.7%	70.1%
Claude-sonnet	89.1%	72.8%
MathFlow-P-7B	**97.2%**	**85.6%**

Table 12: **Generate Config of MathFlow.**

Hyper Parameter	Value
Temperature	0.3
TopP	0.7
TopK	1.0
Repetition Penalty	1.0
Num of Beams	1.0

baselines in both aspects. It achieves a Structured F1 of **97.2%** on EI extraction, demonstrating highly accurate detection of core visual entities, and **85.6%** on RP prediction, indicating strong ability to capture inter-element structural reasoning.

C.5 More Detailed Experiment Results

Generate Config of MathFlow. We first present the generation configuration of MathFlow in Tab. 12. These hyperparameters were carefully tuned to optimize the model's generation performance while maintaining output stability and reliability.

Comprehensive and complete performance comparison across all models. We present comprehensive experimental results through three detailed analyses. First, we evaluate our model's performance across different problem versions in FlowVerse through Tab. 13. This analysis spans six distinct versions incorporating various combinations of textual and visual components: Descriptive Information (DI), Essential Information (EI), Reasoned Property (RP), and Only Question (OQ). The Text Plus Version specifically examines performance without image input. We report both standard accuracy (Acc) and performance with the FlowVerse-CoT-E strategy (CoT-E), highlighting the highest accuracies achieved by closed-source and open-source MLLMs in red and blue respectively.

Full Ablation Analysis of MathFlow on FlowVerse†. Tab. 14 presents a thorough ablation analysis of our perception model on FlowVerse†. This analysis systematically examines the contribution of different components within our perception framework, with bold numbers indicating the best performance across different configurations.

Full Performance Comparison of MLLMs on MathVerse and FlowVerse† Datasets. Tab. 15 provides a comprehensive performance comparison of various MLLMs across both MathVerse and FlowVerse† datasets. Here, FlowVerse† represents the raw version of the dataset, allowing us to evaluate the models' fundamental capabilities without additional enhancements. This comparison offers insights into the relative strengths of different architectures and approaches across diverse mathematical reasoning tasks.

C.6 The more details of the *Image-Only* scenario.

We introduce an *Image-Only* variant (denoted by †) in FlowVerse by superimposing each textual question onto its corresponding diagram, as illustrated in Tab. 16. We then evaluate our MathFlow pipeline on this variant. In the table below, green downward arrows (↓) indicate the performance drop relative to the original multi-input setting. "MathFlow$^*_{4o}$" and "MathFlow$^*_{r1}$" denote MathFlow-P-7B paired with GPT-4o and DeepSeek-R1 inference models, respectively, and T and V abbreviate "Text" and "Vision." These results confirm that MathFlow extends seamlessly to the *Image-Only* scenario with only minor accuracy degradations.

C.7 Analysis of Reasoning Efficiency Across Model Scales

Tab. 17 presents an evaluation of MathFlow built upon Qwen2.5-VL backbones ranging from 3B to 72B parameters, aiming to explore the trade-off between model size, reasoning accuracy, and inference latency on FlowVerse. As expected, both accuracy and inference time increase with model size—for instance, the 3B model achieves 34.2 accuracy in 0.47 seconds, while the 72B variant reaches 58.3 accuracy at the cost of 18.67 seconds per inference. The 7B variant strikes the most favorable balance, attaining 44.8 accuracy in just 2.51 seconds. Notably, the trained variant (MathFlow$_{\text{train}}$) achieves lower latency than the non-fine-tuned counterpart (MathFlow$_{\text{ori}}$), while avoiding redundant reasoning loops that occasionally occur in the latter.

Table 13: **Mathematical Evaluation on Six Problem Versions in FlowVerse**. DI, EI, RP, OQ refer to the textual or visual Descriptive Information, Reasoned Property, Essential Information, Only Question, respectively. The Text Plus Version does not involve image input. "CoT-E" or "Acc" denotes whether to employ the FlowVerse-CoT-E strategy or not. The highest accuracy for each group of MLLMs is marked in **bold**.

Model	All		Text Centric DI+RP+EI+OQ		Text Limited RP+EI+OQ		Text Plus DI+RP+EI+OQ		Vision Dense EI+OQ		Vision Centric RP+EI+OQ		Vision Primary RP+EI+OQ	
	CoT-E	Acc	CoT-E	Acc	CoT-E	Acc	CoT-E	Acc	CoT-E	Acc	CoT-E	Acc	CoT-E	Acc
Open-source MLLMs														
InfiMM-Math-7B	36.5	28.8	43.8	38.1	40.6	36.7	46.1	40.1	28.8	15.4	39.6	30.3	26.1	23.2
InternVL2.5-8B	44.7	41.0	49.2	41.3	40.5	38.4	49.6	42.7	38.4	20.2	41.0	35.9	35.8	33.9
Qwen2.5-VL-7B	53.8	42.2	60.1	52.8	58.9	51.3	62.0	55.0	45.0	31.0	50.8	46.3	48.1	45.3
VLM-R1-7B†	50.7	41.2	59.0	54.2	57.9	49.8	65.5	58.9	36.2	24.5	46.1	37.8	30.6	26.1
Qwen2-VL-72B	52.3	48.6	59.4	47.3	54.3	45.7	63.7	50.0	40.8	25.3	50.9	42.1	47.6	37.0
InternVL2.5-78B	54.7	50.1	66.1	62.7	64.1	60.3	67.8	64.7	48.7	34.3	63.0	58.8	59.6	57.7
MathFlow$^{\star}_{\text{Qwen2.5-VL-7B}}$	57.0	46.0	62.0	53.8	60.8	52.2	64.2	56.0	49.0	39.1	54.5	51.6	52.0	51.5
Math-specialized MLLMs														
MathLLaVA-13B	38.0	29.9	45.1	39.3	44.4	37.4	–	–	36.2	18.6	41.7	35.9	37.0	34.2
MultiMath-7B	44.0	34.2	50.6	44.8	49.9	42.9	–	–	41.7	22.1	47.2	40.4	39.7	38.8
SVE-Math-Qwen2.5-7B	46.0	39.4	53.1	47.3	53.4	45.8	–	–	44.2	28.6	48.9	44.2	45.8	42.0
Closed-source MLLMs														
Qwen-VL-Plus	34.4	27.1	38.1	24.6	35.7	28.6	41.7	35.6	28.5	20.9	32.6	26.0	30.0	22.2
Gemini-Pro	39.0	23.5	44.3	36.5	40.2	33.8	47.2	39.8	32.5	23.3	36.1	27.3	33.8	24.9
Qwen-VL-Max	43.0	36.3	49.8	42.1	46.7	38.3	53.9	51.0	38.6	15.2	42.7	33.2	29.6	27.8
GPT-4o-mini	51.3	44.5	58.7	54.8	58.2	53.2	59.6	55.2	41.1	26.0	57.4	50.1	49.7	47.6
Claude-3.5-Sonnet	56.9	49.6	60.8	52.6	58.7	50.3	64.0	58.3	45.0	25.4	56.5	48.0	48.1	45.2
GPT-4o	55.1	47.8	61.0	56.8	58.7	54.4	62.2	58.2	45.2	30.0	58.6	52.6	54.1	51.0
GPT-4V	56.2	53.4	69.1	57.1	65.0	55.0	72.0	61.4	48.1	30.3	61.8	46.3	42.0	36.7
Gemini-2.5-pro	62.0	55.3	68.3	61.9	66.1	60.8	68.9	64.1	52.1	37.1	65.7	57.9	57.0	54.6
GPT-5	65.8	60.1	74.3	68.1	73.5	66.7	77.0	69.2	53.8	44.7	67.1	61.7	60.3	57.5
MathFlow$^{\star}_{\text{GPT-5}}$	**66.5**	**61.8**	**74.6**	**68.5**	**73.8**	**67.2**	**77.0**	**69.3**	**58.2**	**54.1**	**70.2**	**66.7**	**69.4**	**66.2**

C.8 Qualitative Examples

Figs. 17-19 provide qualitative comparisons between GPT-4V and MathFlow across several geometry problems. These examples illustrate common error types, such as Visual Perception Errors and Reasoning Errors, which are highlighted in the respective model responses.

In the first example, GPT-4V struggles with applying geometric properties correctly, leading to a reasoning mistake while calculating the length of a chord. In contrast, MathFlow arrives at the correct solution by following a systematic application of the Pythagorean theorem and congruent relationships, demonstrating its advantage in maintaining logical accuracy throughout the problem-solving process.

The second example highlights the challenge of dealing with circle theorems. GPT-4V incorrectly deduces the angle at the center due to a misinterpretation of the chord's properties, while MathFlow accurately follows through the geometric relationships to determine the correct angle measure. This example underscores MathFlow's superior ability to navigate complex angle relationships and avoid error propagation in multi-step reasoning.

The third example involves congruent triangles and corresponding parts. GPT-4V makes a visual perception error by misidentifying the corresponding parts of congruent triangles, resulting in an incorrect response. MathFlow, on the other hand, correctly identifies and matches corresponding sides and angles, showcasing its ability to effectively manage both visual and logical components of geometric problems.

As illustrated in Figs. 18–20, end-to-end models often propagate perception errors (such as mislocalized points, incorrect angle estimations, or missed key annotations) into their final answers, resulting in entirely incorrect solutions. **MathFlow adopts a decoupled pipeline with two key advantages: ❶** By isolating the perception stage, MathFlow provides clear and inspectable intermediate outputs that make it straightforward to diagnose and correct vision-based failures without retraining the entire reasoning model. ❷ Once the raw visual content has been translated into structured textual representations, the downstream inference model is relieved of low-level visual decoding tasks. It can apply its full capacity to logical and mathematical reasoning, thus reducing the complexity of the inference model and improving accuracy.

D Limitation and Future Work

While MathFlow and FlowVerse take a step forward in the field of visual mathematical evaluation for MLLMs, it is important to recognize several limitations as follows.

In FlowVerse, we initially categorized the collected mathematical problems based on subjects and subfields that reflect varying degrees of multimodal content. These classification methods

Table 14: **Full Ablation Analysis of MathFlow on FlowVerse†. Bold** numbers indicate the best performance.

Perception Model for EI	Perception Model for RP	Inference Model	COT-E (%)
Qwen2-VL-2B	GPT-4V	GPT-4V	49.0
InternLM-XC2	GPT-4V	GPT-4V	47.1
InfiMM-Math	GPT-4V	GPT-4V	52.6
Qwen2-VL-7B	GPT-4V	GPT-4V	53.3
InternVL2.5-8B	GPT-4V	GPT-4V	53.8
Qwen2.5-VL-7B	GPT-4V	GPT-4V	53.9
MathFlow-P-7B	GPT-4V	GPT-4V	**58.9**
GPT-4V	Qwen2-VL-2B	GPT-4V	49.8
GPT-4V	InternLM-XC2	GPT-4V	50.7
GPT-4V	InfiMM-Math	GPT-4V	54.1
GPT-4V	Qwen2-VL-7B	GPT-4V	52.9
GPT-4V	MathFlow-P-7B	GPT-4V	**57.3**
GPT-4V	GPT-4V	GPT-4V	56.2
MathFlow-P-7B	MathFlow-P-7B	GPT-4V	**59.3**
MathFlow-P-7B	MathFlow-P-7B	GPT-4o-mini	63.8
MathFlow-P-7B	MathFlow-P-7B	Llama3-80B	67.9
MathFlow-P-7B	MathFlow-P-7B	Qwen2.5-72B	69.6
MathFlow-P-7B	MathFlow-P-7B	GPT-4o	69.6
MathFlow-P-7B	MathFlow-P-7B	Gemini 2.5-pro	70.4
MathFlow-P-7B	MathFlow-P-7B	DeepSeek-v3	73.4
MathFlow-P-7B	MathFlow-P-7B	DeepSeek-r1	**75.6**

Table 15: **Full Performance Comparison of MLLMs on MathVerse and FlowVerse† Datasets. FlowVerse†** indicates the raw version of the dataset.

Perception Model	Inference Model	MathVerse	FlowVerse†
InternLM-XC2		25.9	39.6
InfiMM-Math		34.5	47.1
Qwen-VL-MaX		36.2	48.2
Qwen2-VL-72B		38.9	52.3
InternVL-2.5-78B		43.2	54.7
GPT-4V		54.4	56.2
GPT-4o-mini		52.7	59.6
GPT-4o		57.9	62.2
Claude-sonnet-3.5		57.4	64.0
Gemini 2.5-pro		59.9	68.9
MathFlow-P-7B	InternLM-XC2	30.2	43.5
MathFlow-P-7B	InfiMM-Math	38.1	48.9
MathFlow-P-7B	Qwen-VL-MaX	43.3	54.3
MathFlow-P-7B	Qwen2-VL-72B	48.1	58.3
MathFlow-P-7B	InternVL-2.5-78B	56.8	60.1
MathFlow-P-7B	GPT-4V	56.7	59.3
MathFlow-P-7B	GPT-4o-mini	58.9	63.8
MathFlow-P-7B	GPT-4o	59.5	69.6
MathFlow-P-7B	Claude-sonnet-3.5	60.8	68.1
MathFlow-P-7B	Gemini 2.5-pro	62.4	70.4

enable a comprehensive evaluation of MLLM capabilities across different dimensions. However, an additional categorization by difficulty level—similar to the approach used in datasets like MATH (Hendrycks et al., 2021) and WeMath (Qiao et al., 2024), where problems are sorted by complexity—would provide deeper insights into model performance. This extra layer of differentiation could significantly enhance the model's assessment, and we plan to explore this in our future work.

Furthermore, the issues with FlowVerse primarily stem from the fact that they are mainly available in English and Chinese. This limitation restricts the model's applicability to a wider range of languages. By expanding the dataset to include a greater diversity of languages, we could enhance the multilingual capabilities of MLLMs and create benchmarks that are more inclusive for users who speak languages other than English and Chinese.

Within the MathFlow pipeline, although we have trained the perception model using the RP captioning task, the resulting reasoning capabilities remain relatively basic, making it challenging for the model to handle more complex reasoning tasks effectively. On the other hand, with the emergence of models (OpenAI, 2023c; Zhang et al., 2024b; Xu et al., 2024) like GPT-o1 (OpenAI, 2023c) and LLaMA-o1 (Zhang et al., 2024b), which demonstrate enhanced capabilities, we see an opportunity for further advancement. Consequently, in future work, we plan to build upon these advanced models to improve the reasoning capabilities of MathFlow.

Table 16: **The more details of the *Image-Only* scenario.**

Model	**T-Centric**†	**T-Limited**†	**V-Dense**†	**V-Centric**†	**V-Primary**†
MathFlow$^{\star}_{GpT-4o}$	68.5 (0.61 ↓)	67.4 (0.80 ↓)	52.8 (1.37 ↓)	55.4 (1.03 ↓)	54.7 (1.55 ↓)
MathFlow$^{\star}_{DeepSekk-r1}$	72.1 (0.52 ↓)	70.4 (0.62 ↓)	57.3 (0.97 ↓)	65.7 (1.14 ↓)	62.6 (1.39 ↓)

Table 17: **Reasoning efficiency of MathFlow built on Qwen2.5-VL models of varying sizes.** The trained variant MathFlowF$_{\text{train}}$) demonstrates faster inference than the non-fine-tuned version (MathFlow$_{\text{ori}}$), especially on the 7B model, which offers the best balance between speed and performance.

Model	**Size=3B**	**Size=7B**	**Size=32B**	**Size=72B**
End2End (CoT)	34.2 (0.47s)	44.8 (2.51s)	51.1 (6.45s)	58.3 (18.67s)
MathFlow$_{\text{ori}}$ (CoT)	37.7 (0.68s)	46.7 (3.46s)	51.1 (9.13s)	59.2 (26.58s)
MathFlow$_{\text{train}}$ (CoT)	48.0 (0.63s)	51.7 (3.19s)	-	-

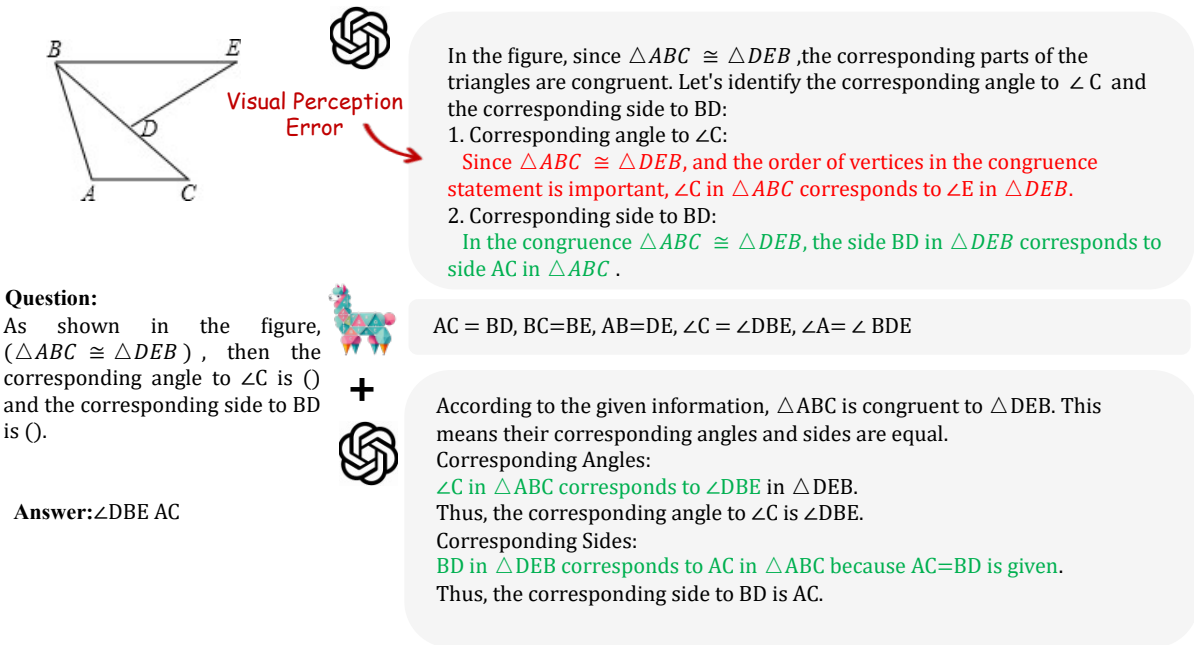
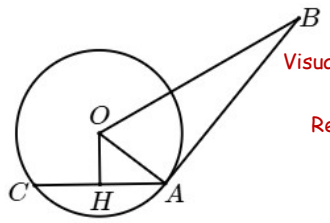


Figure 17: **Response Comparison of GPT-4V and MathFlow**$^{\star}{}_{\text{GPT-4V}}$



Question:
AB is a tangent to circle O, A is the point of tangency, AC is a chord of circle O, a perpendicular from O to AC intersects at H. If OH = 3, AB = 12, BO = 13, then the length of chord AC is ( )

Answer: 8

Visual Perception Error

Reasoning Error

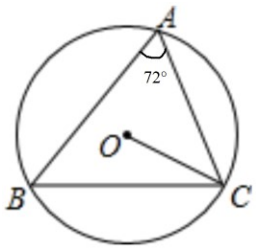
To find the length of chord AC, we apply the Pythagorean theorem in the right triangle OAB, where O is the center of the circle, A and B are points on the circle, and H is the foot of the perpendicular from O to AB. Given that OH = 3 (radius), AB = 12, and BO = 13, we check if the triangle is right-angled using the Pythagorean theorem: ($OH^2 + AH^2 = BO^2$). This yields ($3^2 + AH^2 = 13^2$), resulting in ($AH = 4\sqrt{10}$). Since H bisects AB, the length of AC is half of AB, so AC = 6 . The problem's choices (A, B, C, D) do not match this result, indicating a potential error in the question setup. Therefore, the length of chord AC is 6.

∠OAB=90°, AO = $\sqrt{OB^2 - AB^2}$ = 5, AC = 2AH

+

From the given information, we can determine the length of chord AC in the circle. 1. We know that $\triangle OAB$ is a right triangle with ∠OAB = 90° , OB = 13 (radius of the circle), and AB = 12 . Using the Pythagorean theorem, ($AO = \sqrt{OB^2 - AB^2} = \sqrt{13^2 - 12^2} = \sqrt{169 - 144} = \sqrt{25} = 5$). 2. It is given that AC = 2AH, $AH = \sqrt{AO^2 - HO^2} = \sqrt{5^2 - 3^2} = \sqrt{25 - 9} = \sqrt{16} = 4$. Therefore, $AC = 2 \times 4 = 8$. Thus, the length of chord AC is 8. The answer to the question is 8

Figure 18: **Response Comparison of GPT-4V and MathFlow**$^{\star}{}_{\mathrm{GPT-4V}}$



Question:
As shown in the figure, ⊙O is the circumcircle of △ABC, then the degree measure of ∠BCO is

Answer: 18

Reasoning Error

Given that ∠A = 72° and ⊙O is the circumcircle of triangle ABC, we need to find the measure of ∠BCO. In a triangle, the angle at the center of the circumcircle (O) that subtends a chord (BC in this case) is twice the angle at the circumference subtended by the same chord. Therefore, if ∠A subtends chord BC at the circumference, then ∠BOC at the center subtending the same chord BC will be twice ∠A. ∠BOC = 2 × ∠A = 2 × 72° = 144° Since ∠BOC is the full angle between line segments BO and CO, and ∠BCO is half of this angle (as ∠BCO and ∠B'CO are equal, where B' is the point where line BO intersects the circle at the other side), we have: ∠BCO = 1/2 × ∠BOC = 1/2 × 144° = 72° Therefore, the measure of ∠BCO is 72°.

∠BOC=2∠A=144°, ∠CBO=∠BCO, ∠BOC+∠CBO+∠BCO=180°

+

Given that ∠BOC = 144° and ∠CBO = ∠BCO, and knowing that the sum of angles in a triangle is 180°, we can set up the equation: ∠BOC + ∠CBO + ∠BCO = 180° Substituting the given values: 144° + ∠CBO + ∠BCO = 180° Since ∠CBO = ∠BCO, let x be the measure of each of these angles. Then: 144° + x + x = 180° 144° + 2x = 180° 2x = 180° - 144° 2x = 36° x = 18° Thus, ∠BCO = 18°. Answer: 18°

Figure 19: **Response Comparison of GPT-4V and MathFlow**$^{\star}{}_{\mathrm{GPT-4V}}$