VISIOMATH: BENCHMARKING FIGURE-BASED MATHEMATICAL REASONING IN LMMS

Can Li **Ying Liu** **Ting Zhang** **Mei Wang** **Hua Huang**
Beijing Normal University

ABSTRACT

Large Multimodal Models have achieved remarkable progress in integrating vision and language, enabling strong performance across perception, reasoning, and domain-specific tasks. However, their capacity to reason over multiple, visually similar inputs remains insufficiently explored. Such fine-grained comparative reasoning is central to real-world tasks, especially in mathematics and education, where learners must often distinguish between nearly identical diagrams to identify correct solutions. To address this gap, we present VisioMath, a curated benchmark of 1,800 high-quality K–12 mathematics problems in which all candidate answers are diagrams with subtle visual similarities. A comprehensive evaluation of state-of-the-art LMMs, covering both leading closed-source systems and widely adopted open-source models, reveals a consistent decline in accuracy as inter-image similarity increases. Analysis indicates that the dominant failure mode stems from image–text misalignment: rather than grounding reasoning in textual cues, models often resort to shallow positional heuristics, resulting in systematic errors. We further explore three alignment-oriented strategies, spanning training-free approaches and finetuning, and achieve substantial accuracy gains. We hope that VisioMath will serve as a rigorous benchmark and catalyst for developing LMMs toward deeper diagram understanding, precise comparative reasoning, and grounded multi-image–text integration.

1 INTRODUCTION

In recent years, Large Multimodal Models (LMMs) (Chen et al., 2025; OpenAI, 2024a; Team, 2024a; Wang et al., 2024c; Wu et al., 2024b) have achieved remarkable success across various multimodal tasks. This surge in capabilities is largely attributed to the availability of massive, high-quality vision-and-language datasets (Chen et al., 2023; He et al., 2023; Kuznetsova et al., 2020; Singla et al., 2024), which enable the training of increasingly capable models. By jointly modeling image and text modalities, LMMs enable seamless cross-modal reasoning, allowing for the interpretation of complex visual scenes in natural language and vice versa. This integration not only enhances basic perceptual capabilities but also supports high-level cognitive tasks such as visual recognition (Chen et al., 2024b; Huang et al., 2024; Wang et al., 2024d), logical reasoning (Wang et al., 2024e; Wu et al., 2024a; Xiao et al., 2024), and context understanding (Zhang et al., 2024a).

With the rapid development of LMMs, designing holistic benchmarks is essential for systematically investigating the capabilities and limitations of these models. Numerous evaluation benchmarks have been proposed, targeting different aspects of LMM performance, including perception, reasoning, domain-specific tasks, hallucination, and multimodal integration (Huang & Zhang, 2024; Li et al., 2024c). Among these, multimodal reasoning ability, particularly mathematical reasoning that requires integrating visual and textual information, has always been a central focus. This form of reasoning presents distinct challenges, requiring not only the understanding of mathematical semantics in text but also the accurate interpretation and synthesis of visual representations.

To evaluate multimodal reasoning capabilities, various multimodal mathematical reasoning benchmarks have been introduced (Lu et al., 2024; Zhang et al., 2024b; Wang et al., 2024b). These benchmarks can be broadly divided into two categories. The first involves single-image scenario, where each problem is paired with a single diagram that supplements the text. While effective for assessing basic multimodal understanding, these setups are limited in capturing the complexity of

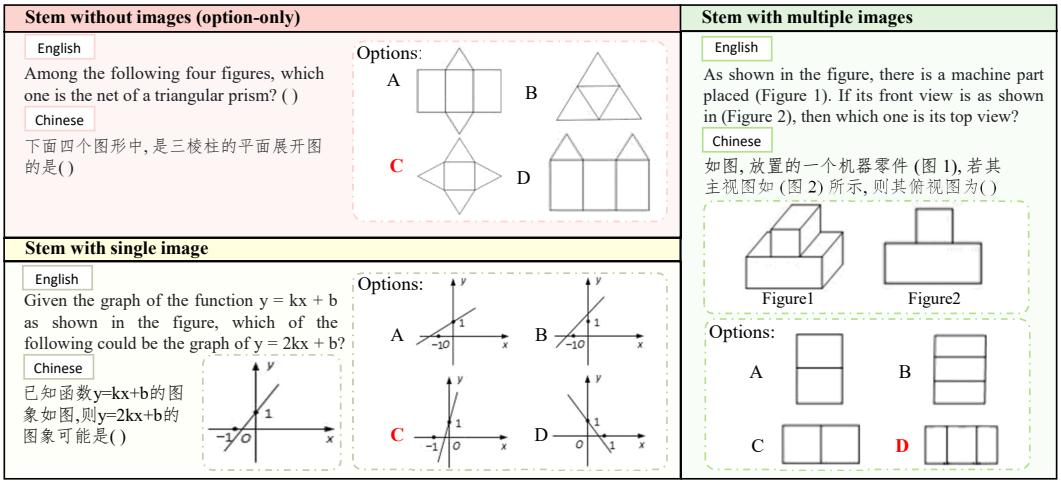


Figure 1: Illustrating examples in our VisioMath dataset, in which data samples consist of visual answer options exhibiting high visual similarity, and the stem may appear with or without images.

real-world visual reasoning, as a single image often lacks the richness and inter-image dependencies needed for higher-order comprehension. In response, recent studies have shifted toward the second category: multi-image scenario. These tasks require reasoning across problems with multiple visual inputs. This paper also investigates on multi-image scenario with a particular emphasis on a specific and underexplored setting: reasoning over multiple highly similar images.

In this paper, we examine a distinct class of multimodal benchmarks in which all answer choices are presented as images. Our motivation arises from the observation that many real-world mathematical problems, especially in educational settings, present options as diagrams (e.g. geometric figures). Addressing such problems involves more than visual recognition; it necessitates comparison of visually similar structures and reasoning about subtle symbolic differences. While recent benchmarks such as CMM-Math-test (Liu et al., 2024b), MathVerse-mv (Li et al., 2024b), and MV-Math (Wang et al., 2025) have advanced the evaluation of multimodal reasoning by introducing multi-image questions, they often overlook a crucial aspect where reasoning must be grounded in perceptually similar visual features. Our work aims to address this gap and thereby provide an evaluation perspective that specifically targets LMMs' reasoning across closely resembling images.

To achieve that, we introduce VisioMath, a novel benchmark comprising 1,800 meticulously curated, high-quality mathematics problems. The dataset spans a broad spectrum of K–12 mathematics topics, including geometry, algebraic visualizations, numerical comparisons, and functional pattern recognition, thereby capturing the diversity of real-world curricula. Each problem features diagrammatic answer options, with approximately 50% also incorporating at least one image in the question stem to provide essential visual context. To ensure accuracy and reliability, each question has been independently annotated and cross-validated by at least two expert annotators. To reduce answer-choice bias in LMMs, we enforce a uniform distribution across the four multiple-choice options (A, B, C, D). As shown in Figure 1, each answer option is a distinct diagram differing subtly from the others, requiring fine-grained visual discrimination.

We conduct a comprehensive evaluation on the VisioMath benchmark. Our study encompassed a diverse set of LMMs across various model families and scales, including state-of-the-art closed-source models such as GPT-4.1 (OpenAI, 2025) and Gemini2.5 Pro (Comanici et al., 2025), as well as prominent open-source models like Qwen2.5-VL (Bai et al., 2025). These models span different input paradigms, with some restricted to single-image inputs and others capable of processing multiple images simultaneously. We perform a detailed error analysis and find that image–text misalignment accounts for the largest proportion. This highlight a fundamental overlooked limitation in current LMMs: their inability to reliably establish fine-grained correspondences between multiple images and distinct textual inputs. In tasks such as figure–option problems, where each figure must be uniquely paired with a specific textual option, LMMs often fail to preserve these one-to-one mappings. This weakness indicates that, although LMMs excel in single-image reasoning and

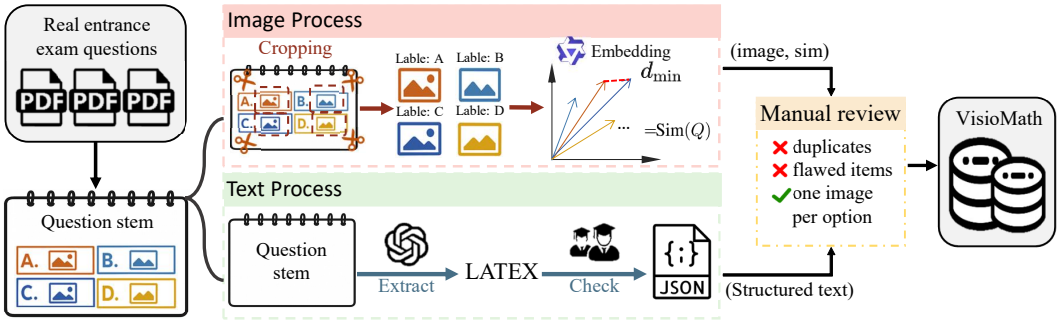


Figure 2: Data processing pipeline of VISIOMATH, including text extraction and verification, image cropping, and integration of visual similarity information to construct the final dataset.

holistic multimodal understanding, they remain inadequate when tasks demand precise cross-modal alignment across multiple visual–text pairs.

We further explore three complementary strategies aimed at mitigating image–text misalignment and enhancing multi-image reasoning: consolidating multiple images into a single layout, establishing explicit visual–textual anchors, and fine-tuning with an alignment-oriented multi-image chain-of-thought dataset. Notably, such limited Chain-of-Thought(CoT) fine-tuning data yields a substantial accuracy gain (+12.6%), illustrating the critical role of explicit visual–textual alignment in enabling effective multi-image reasoning. We hope our work will motivate more systematic exploration of methods for enhancing multi-image–text alignment in complex reasoning tasks.

In summary, our key contributions are:

- **VisioMath Benchmark.** We introduce VisioMath, the first benchmark specifically designed for visual-option mathematical reasoning. It bridges the gap between traditional multimodal visual question-answering benchmarks, providing a rigorous testbed for evaluating LMMs' diagram understanding and fine-grained visual reasoning.
- **Comprehensive Evaluation.** We systematically evaluate a wide range of state-of-the-art LMMs, including GPT-4.1 and Gemini2.5 pro, and reveal that even top-performing models struggle with reasoning over visually similar answer options, highlighting a critical limitation when dealing with complex reasoning requiring multi-image-text alignment.
- **Analytical Strategies.** We perform detailed error analyses to identify key failure modes, design controlled experiments to validate the critical limitation, and introduce alignment-focused strategies that substantially improve figure-based reasoning performance.

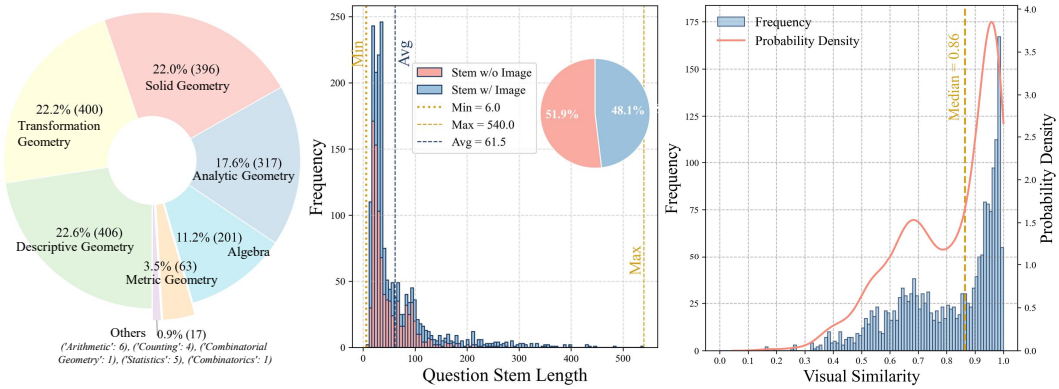
2 VISIOMATH

Motivation. In mathematics education, multiple-choice questions with diagrammatic answer options are pervasive. These diagrams often exhibit high visual similarity, differing only in subtle geometric structures or functional curves. Humans can reliably leverage such fine-grained differences through prior knowledge and structured reasoning. In contrast, LMMs typically rely on superficial embedding similarity, making it difficult to discriminate between nearly identical options.

Routine for students, this setting remains unexpectedly challenging for LMMs. As illustrated in Figure 1, the four candidate diagrams share almost identical visual styles, yet solving the problem requires aligning the textual description with precise visual interpretation. To capture this ubiquitous but underexplored scenario, we introduce VisioMath, a benchmark explicitly designed to evaluate LMMs' reasoning ability over multiple highly similar visual options in mathematics.

2.1 BENCHMARK CONSTRUCTION

Building on the motivation introduced above, VisioMath is constructed to faithfully reproduce exam-like scenarios where reasoning hinges on subtle visual distinctions. To this end, during construction



(a) Distribution of subjects. (b) Distribution of question stem length. (c) Distribution of visual similarity.

Figure 3: Detailed statistics of the VisioMath dataset. The figure shows distributions of (a) subject, (b) question stem length and (c) visual similarity, highlighting both textual and visual characteristics.

we follow three design principles, *representativity, reliability, and high visual similarity*. The overall construction pipeline is illustrated in Figure 2.

Representativity. VisioMath contains 1,800 multiple-choice questions with 8,070 diagrammatic options, collected from Chinese high school and college entrance examinations administered between 2002 and 2023. Using real exam items ensures external validity: the benchmark directly reflects the types of problems students actually face, and its results can generalize to real educational scenarios. Each problem is paired with option diagrams as well as stem diagrams (average 4.48), and we intentionally balanced the correct answer distribution across A–D (24–26% each) to eliminate positional bias. The question stem length average 61.5 tokens, reflecting moderate linguistic complexity as shown in Figure 3 (b). We also present the distribution of subject areas across the dataset in Figure 3 (a), offering an integrated overview of the benchmark's coverage.

Reliability. To ensure that evaluation results reflect genuine reasoning ability rather than spurious cues, we standardize and curate all samples. Question texts were digitized into a consistent JSON format, where mathematical expressions were transcribed into LaTeX to guarantee uniform parsing. Answer diagrams were carefully cropped from PDFs to enforce a strict one image per option rule, preventing layout or formatting artifacts from providing shortcuts. Finally, all items underwent manual review to eliminate duplicates, low-quality images, and conceptually flawed questions. These steps establish a dataset that is reliable to faithfully evaluate the visual reasoning ability of LMMs.

High Visual Similarity. A distinctive property of VisioMath lies in its *systematic quantification of visual similarity* among answer options. For each question Q, option images x_i are encoded using the Qwen multimodal-embedding-v1 model, and the question-level visual similarity is then defined as the minimum pairwise cosine similarity across all encoded images:

$$\text{Sim}(Q) = \min_{i \neq j} \cos\big(f(x_i), f(x_j)\big), \tag{1}$$

where $f(\cdot)$ denotes the image embedding encoder.

As illustrated in Figure 3 (c), a large proportion of VisioMath problems contain *highly similar options*, creating fine-grained distinctions that are especially challenging for LMMs. Importantly, we preserve the full spectrum of similarity levels rather than filtering out low-similarity cases, so that performance can be systematically compared under different similarity regimes.

2.2 Benchmark Analysis

Unique Challenges. VisioMath introduces a set of unique challenges that distinguish it from existing multimodal benchmarks. Unlike conventional tasks that pair a single image with text, VisioMath requires reasoning across multiple diagrammatic options simultaneously, transforming the problem into one of comparative visual reasoning that mirrors authentic exam scenarios. Moreover, the benchmark faithfully preserves the presence of highly similar distractors, where candidate dia-

Table 1: Comparison between VisioMath and existing evaluation datasets. Here, EN and CN denote English and Chinese, respectively; FO refers to figure-based options; and AvgImg indicates the average number of images for each problem.

Datasets	Multi-image problem	Language	#Problems (FO)	#Problems	#Images	AvgImg
We-Math (Qiao et al., 2024)	✘	EN	–	6500	6500	1
MMMU-Math (Yue et al., 2024)	✘	EN	–	540	540	1
Math-Vista (Lu et al., 2024)	✘	EN	–	6141	6141	1
Math-Verse (Zhang et al., 2024b)	✘	EN	–	2612	2612	1
Math-Vision (Wang et al., 2024b)	✘	EN	–	3040	3040	1
MM-Math (Sun et al., 2024)	✘	EN	–	5,929	5,929	1
CMMU-MATH (He et al., 2024)	✘	CN	–	778	778	1
MathExplain (Park et al., 2025)	✘	EN	–	997	997	1
MathGlance (Sun et al., 2025)	✘	EN	–	1,609	1,609	1
Gaokao-MM-Math (Zong & Qiu, 2024)	✔	CN	17	80	142	1.78
CMM-Math-test (Liu et al., 2024b)	✔	CN	245	5821	3794	2.26
MathVerse-mv (Li et al., 2024b)	✔	EN	0	788	6304	8
MV-Math (Wang et al., 2025)	✔	CN,EN	595	2009	6061	3.02
VisioMath (Ours)	✔	CN,EN	1800	1800	8070	4.48

grams differ only in subtle geometric or symbolic details, thereby testing models' capacity for fine-grained perceptual discrimination. Finally, effective problem solving in VisioMath demands precise text–visual alignment, as models must ground linguistic conditions such as symmetry, monotonicity, or functional transformations in the correct visual choice. Collectively, these characteristics elevate VisioMath from simple image recognition to a rigorous evaluation of figure-based visual reasoning.

Benchmark Comparison. We compare VisioMath with prior multimodal math benchmarks in Table 1. Most existing datasets adopt a single-image setting with textual answer options (e.g., We-Math, MMMU, Math-Vista, Math-Verse, Math-Vision). Multi-image formats are rare, and when present, figure-based options are either absent or inconsistently represented. For instance, MathVerse-mv includes multiple images but no visual answer options. CMM-Math-test and MV-Math contain some image-based options, yet many are embedded in composite layouts rather than provided as independent visual elements. VisioMath, in contrast, explicitly structures answer options as collections of distinct and semantically meaningful images, thereby supporting a more nuanced evaluation of fine-grained visual mathematical reasoning.

3 EXPERIMENT

Setup. To comprehensively evaluate the performance of LMMs in handling complex visual inputs, we select a diverse set of models across different accessibility types and input configurations. Specifically, we include closed-source LMMs representing the current state-of-the-art in commercial multimodal systems. In addition, we conduct experiments on open-source LMMs that explicitly support multi-image inputs with various model sizes. This broad coverage ensures a representative analysis across model capacities and architectures. Moreover, to evaluate the adaptability of models not originally designed for multi-image processing, we also evaluate several widely-used single-image input LMMs. For these models, we implement a composite image concatenation strategy, in which all images associated with a given question were merged into a single composite one. All LMMs are evaluated under a zero-shot setting to ensure a fair and consistent comparison of their generalization capabilities. More details are provided in Appendix A.

3.1 RESULTS

Table 2 reports the comparative performance of various LMMs on VisioMath benchmark, with results disaggregated by the ground-truth (GT) answer position (A–D). The evaluation considers two distinct conditions: (i) question stems presented without images and (ii) question stems accompanied by images. For each condition, we provide both average accuracy and per-option performance. Table 3 further details the accuracy of LMMs on subsets of the dataset stratified by image similarity levels. The dataset is divided into quartiles based on the degree of visual similarity between images within each question, and model performance is reported separately for each quartile. This analysis aims to evaluate models' fine-grained reasoning capabilities under varying visual similarities.

Based on these results, we have following observations.

Table 2: Performance comparison on VisioMath with results categorized based on GT position.

Models \GT position	Avg	Question stem w/o images					Question stem with images				
		Avg	A	B	C	D	Avg	A	B	C	D
Random	25.6	25.4	24.0	25.6	23.0	28.6	26.0	22.8	27.6	28.4	25.6
Closed-source LMMs											
QwenVL-plus (Bai et al., 2023)	32.9	39.1	27.0	59.9	43.4	25.5	26.3	7.5	26.2	34.8	40.1
QwenVL-max (Bai et al., 2023)	44.1	53.4	35.2	62.6	62.5	50.2	34.1	31.1	34.1	32.8	38.6
GPT-4o (OpenAI, 2024a)	45.9	54.7	55.6	56.4	54.7	52.5	36.5	47.3	30.4	36.3	30.4
GPT4.1 (OpenAI, 2025)	52.6	61.6	72.4	59.9	60.2	56.1	42.8	54.8	39.3	43.7	31.9
Gemini2-flash-thinking (DeepMind, 2025b)	53.2	61.2	80.6	59.9	58.6	50.3	44.6	57.3	43.0	43.1	32.9
GLM-4.5V (Team et al., 2025)	53.7	69.1	71.9	75.8	68.4	61.2	37.2	46.5	42.5	31.4	26.6
Gemini2-flash (DeepMind, 2025a)	55.5	65.1	78.1	59.9	65.2	59.6	45.2	57.7	34.5	38.7	47.8
Doubao-1.5-Vision-pro (Team, 2025a)	66.3	75.6	78.6	78.0	76.6	70.2	56.4	73.4	55.6	48.0	45.4
Seed1.6-Thinking (ByteDance, 2024)	72.3	85.7	**90.3**	**87.2**	82.4	83.9	58.0	71.8	53.7	44.6	59.4
Gemini 2.5 Pro (Comanici et al., 2025)	**80.9**	**86.2**	89.2	84.6	**85.2**	**86.3**	**75.2**	**78.8**	**77.6**	**75.0**	**68.6**
Open-source LMMs (multi-image input)											
InternVL2.5-2B (Chen et al., 2024a)	24.6	27.1	12.8	25.5	36.3	30.2	21.9	10.3	26.2	38.2	15.0
Qwen2.5-VL-3B-instruct (Bai et al., 2025)	25.4	26.1	51.0	40.5	14.5	5.9	24.7	18.3	**70.1**	5.4	4.3
R1-Onevison-7B (Yang et al., 2025)	29.6	35.0	38.8	37.4	34.8	30.2	23.7	22.0	32.2	28.9	11.6
QvQ-72B-Preview (Team, 2024b)	30.9	36.2	45.4	36.1	44.0	31.4	25.3	31.5	28.0	26.0	14.5
Qwen2-VL-72B-instruct (Wang et al., 2024c)	31.7	38.2	15.8	29.5	**78.1**	23.1	24.5	2.5	9.8	**83.8**	6.8
Qwen2.5-VL-7B-instruct (Bai et al., 2025)	32.7	39.5	30.1	58.1	39.8	29.8	25.3	8.7	28.5	32.4	34.3
Gemma3-27B (Team, 2025b)	35.3	43.7	**67.9**	40.1	33.6	38.4	26.2	**40.2**	24.8	12.3	25.1
Vision-R1-7B (Huang et al., 2025)	36.7	43.7	47.4	57.3	38.7	33.7	29.2	24.5	52.3	29.4	10.6
Qwen2.5-VL-72B-instruct (Bai et al., 2025)	**43.7**	**53.5**	36.2	**63.9**	61.3	**49.8**	**33.0**	29.9	37.8	29.9	**35.2**
Open-source LMMs (single-image input)											
MiniCPM-v2.5 (Abdin et al., 2024)	21.0	21.7	28.1	13.2	12.1	34.1	20.2	28.2	15.4	6.4	**29.5**
GLM4V-9B (GLM, 2024)	23.9	**25.6**	19.4	**31.7**	**31.6**	18.8	22.2	10.3	**33.2**	**26.0**	20.8
LLaVA-v1.6-vicuna-13B (Liu et al., 2024a)	**24.4**	23.0	**50.5**	2.2	5.1	**38.4**	**26.0**	**66.4**	0.0	2.9	28.5

Observation 1 (*Limited performance of single-image LMMs in multi-image reasoning tasks*). To evaluate the capability of single-image LMMs in multi-image reasoning tasks, we employ a simple strategy: concatenating multiple images into a single composite and applying single-image LMMs for reasoning. Despite its straightforwardness, this approach exposes critical limitations. Among the models evaluated, the best performer, LLaVA-v1.6-vicuna-13B achieves only 24.4% accuracy, on par with the naive baseline, namely random guessing. These results underscore a fundamental limitation of single-image LMMs in multi-image contexts:they fail to effectively model relational information across distinct visual inputs. This highlights the need for architectures that explicitly support cross-image representation learning and comparative reasoning.

Observation 2 (*Question stems containing images pose greater challenges for LMMs*). As shown in Table 2, most LMMs demonstrate noticeably lower performance on questions whose stems include images compared to those with text-only stems, which is a trend consistent across nearly all positions. This observation suggests that the inclusion of images in the question stem significantly increases the complexity of the visual reasoning task. Specifically, when both the stem and the options involve visual content, LMMs are required to process and integrate multiple sources of visual information, which likely imposes a higher cognitive load on the model. This indicates that current LMMs still struggle with multi-image reasoning scenarios and highlights a potential bottleneck in their capacity for holistic visual understanding.

Observation 3 (*Performance degrades under high visual similarity*). LMMs exhibit performance degradation on tasks involving high inter-image similarity, as shown in Table 3. For instance, Doubao-1.5-Vision-Pro achieves 74.9% accuracy in the quartile with the lowest similarity, but this drops to 62.0% in the highest-similarity quartile, a 12.9% decline. This performance gap stems from the increased demands for fine-grained cross-image associative reasoning, which current LMMs insufficiently support due to limited visual granularity and reasoning capabilities. Notably, LMMs exhibit strong performance correlations across similarity quartiles: models performing well in low-similarity scenarios tend to retain relative strength under high similarity.

3.2 Analysis

Error Categorization. We conduct a systematic error analysis of GLM4.5V to better understand the limitations of LMMs on VisioMath. From the model outputs, we randomly sample 50 erroneous cases and manually inspect their characteristics, and we categorize the errors into five types,

Table 3: Performance comparison on VisioMath with results categorized based on image similarity.

Models \Image similarity	Avg	[0.16,0.68]	(0.68,0.87]	(0.87,0.96]	(0.96,1]
Random	25.6	23.6	24.4	27.8	27.1
Closed-source LMMs					
QwenVL-plus (Bai et al., 2023)	32.9	33.3	37.8	32.4	28.2
QwenVL-max (Bai et al., 2023)	44.1	47.3	50.2	41.3	37.6
GPT-4o (OpenAI, 2024a)	45.9	53.8	50.9	40.0	39.1
GPT-4.1 (OpenAI, 2025)	52.6	65.8	56.4	42.9	45.1
Gemini2-flash-thinking (DeepMind, 2025b)	53.2	63.6	58.9	48.2	42.2
GLM-4.5V (Team et al., 2025)	53.7	68.7	59.3	44.2	44.7
Gemini2-flash (DeepMind, 2025a)	55.5	66.7	59.8	49.3	46.2
Doubao-1.5-Vision-pro (Team, 2025a)	66.3	74.9	68.2	60.2	62.0
Seed1.6-Thinking (ByteDance, 2024)	72.3	82.4	74.2	66.2	66.4
Gemini 2.5 Pro (Comanici et al., 2025)	**80.9**	**86.2**	**83.8**	**76.7**	**76.9**
Open-source LMMs (multi-image input)					
InternVL2.5-2B (Chen et al., 2024a)	24.6	24.2	28.9	22.7	22.7
Qwen2.5-VL-3B-instruct (Bai et al., 2025)	25.4	26.7	27.6	24.4	22.9
R1-Onevison-7B (Yang et al., 2025)	29.6	21.9	32.2	28.9	11.6
QvQ-72B-Preview (Team, 2024b)	30.9	37.3	38.0	25.3	23.1
Qwen2-VL-72B-instruct (Wang et al., 2024c)	31.7	35.5	37.8	26.0	27.1
Qwen2.5-VL-7B-instruct (Bai et al., 2025)	32.7	33.6	37.8	29.8	29.6
Gemma3-27B (Team, 2025b)	35.3	43.3	41.2	29.6	26.4
Vision-R1-7B (Huang et al., 2025)	36.7	46.7	38.9	30.4	30.9
Qwen2.5-VL-72B-instruct (Bai et al., 2025)	**43.7**	**47.1**	**50.8**	**38.0**	**38.7**
Open-source LMMs (single-image input)					
MiniCPM-v2.5 (Abdin et al., 2024)	21.0	21.7	21.3	20.6	20.2
GLM4V-9B (GLM, 2024)	23.9	**26.7**	23.5	23.3	**22.0**
LLaVA-v1.6-vicuna-13B (Liu et al., 2024a)	**24.4**	24.0	**26.0**	**26.0**	21.8

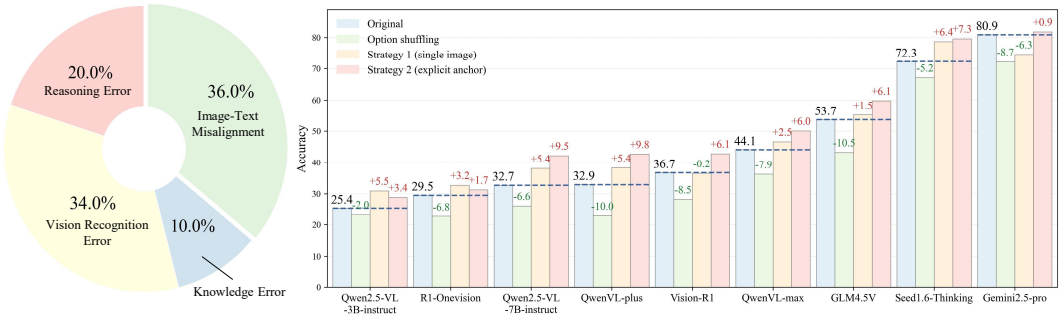
with their proportions illustrated in Figure 4 (a). Among the identified categories, *image–text misalignment* account for the largest shares, representing 36% of the errors. Compared to single-image datasets such as MATH-Vision, these proportions are significantly higher. This finding highlights that reasoning over multiple visual contexts introduces substantial challenges, particularly in maintaining consistent semantic alignment across both images and text.

Effect of Option Shuffling. The image–text misalignment errors suggest that current LMMs rely heavily on heuristic positional correspondences between options and images. To investigate this, we conducted a controlled shuffling experiment: the image order was kept unchanged, while the textual references to the options were permuted. For example, the original prompt "The last four pictures are respectively the pictures for options A, B, C, and D" was modified to "The last four pictures are respectively the pictures for options B, C, D, and A," with the ground-truth answers adjusted accordingly. By keeping the image order constant, we isolate the effect of image order on performance. Results shown in Figure 4 (b) suggest a consistent and clear decline under this manipulation. For instance, Gemini 2.5 Pro's accuracy dropped from 80.9% to 72.2% (–8.7%). These findings indicate that existing LMMs struggle to robustly capture and align semantic correspondences between textual options and visual content, highlighting the need for improved cross-modal alignment mechanisms in multi-image reasoning tasks.

3.3 Strategies for Performance Enhancement

Building on the above analysis of LMM limitations, we explore practical strategies to improve multi-image reasoning performance on VisioMath. These strategies fall into two categories: training-free techniques that leverage structural or labeling cues, and a training-based approach that incorporates specialized multi-image reasoning data. Collectively, they demonstrate the potential to mitigate vision–text misalignment and enhance cross-figure reasoning.

Strategy 1 (*Consolidated single image layout*). We first examine whether providing all visual information in a single spatial layout improves reasoning. Specifically, option images and stem images are concatenated into a composite image. As shown in Figure 4 (b), this structural simplification consistently improves performance, suggesting that LMMs struggle to distribute attention effectively across multiple independent images. For instance, Seed1.6-Thinking achieves an accuracy increase from 72.3% to 78.7% (+6.4%) under this setup. The results indicate that co-locating visual information helps LMMs reason more effectively over multiple images.



(a) Distribution of error types.

(b) Performance of LMMs across Original, Shuffling, and Strategy Variants.

Figure 4: Illustrating error type distribution and the impact of input data structure on performance.

Table 4: The effect of strategy 3, using alignment-oriented multi-image chain-of-thought fine-tuning.

Model	Original	Shuffling	Strategy 1	Strategy 2	Strategy 3
Qwen2.5-VL-3B-instruct (Bai et al., 2025)	25.4	23.4 (-2.0)	30.9 (+5.5)	28.8 (+3.4)	38.0 (+12.6)

Strategy 2 (*Explicit visual–textual anchors*). In this strategy, each image is directly associated with its corresponding textual label, either through overlaid or embedded annotations. This experimental setting is designed to evaluate whether establishing explicit visual–textual correspondences can enhance disambiguation and support more accurate decision-making. Empirical results shown in Figure 4 (b) demonstrate that this approach yields notable performance gains: for instance, QwenVL-plus improves from 32.9% to 42.7% (+9.8%), whereas Gemini 2.5 Pro shows a smaller but measurable gain of +0.9%. These results indicate that current LMMs continue to struggle with robustly binding textual content to the corresponding visual elements. Importantly, the findings highlight that carefully designed visual–textual anchors can effectively mitigate misalignment errors, offering a practical pathway to improve multimodal reasoning performance.

Strategy 3 (*Alignment-oriented multi-image chain-of-thought training*). To further enhance reasoning performance, we develop a specialized multi-image CoT dataset explicitly aimed at improving visual–textual alignment across multiple diagrams. Starting from 1,072 multi-image problems collected online, we first employ QwenVL-Max with an image-caption–style prompt to generate preliminary reasoning paths that describe each diagram individually, ensuring localized alignment between visual elements and textual explanations. These initial outputs are then refined by DeepSeek V3.1 through a CoT Data Generation Prompt, which enforces step-by-step integration of per-image descriptions into a globally coherent reasoning trajectory, tightly binding visual observations to textual inferences. To guarantee reliability, only samples yielding correct final answers are retained, resulting in 500 high-quality multi-image CoT exemplars with explicit visual–textual anchors. We fine-tuning Qwen2.5-VL-3B on this dataset, and the results in Table 4 show that accuracy increases from 25.4% to 38.0% (+12.6%), surpassing R1-OneVision-7B (29.5%) and Vision-R1 (36.7%), despite using only a small set of CoT data. These results highlight that current models are severely constrained by the scarcity of alignment-oriented multi-image CoT training data, and that targeted augmentation with explicit alignment signals can substantially boost figure-based reasoning.

4 Related Work

Large Multimodal Models. Recently, LMMs have achieved remarkable progress in visual understanding (OpenAI, 2023; 2024b). Early LMMs were predominantly trained on single-image visual question answering datasets, limiting their ability to reason across multiple images. Recent progress has alleviated this limitation through interleaved image-text pretraining on large-scale corpora such as MMC4 (Zhu et al., 2023) and Omnicorpus (Li et al., 2024d). Additionally, instruction tuning using datasets like Mantis-Instruct (Jiang et al., 2024) has enhanced models' alignment with human instructions in multi-image scenario. As a result, advanced LMMs like GPT-4o (OpenAI, 2024a) and Gemini2.5 Pro (Comanici et al., 2025) exhibit robust capabilities in tasks involving image count-

ing, comparison, and comprehension in diverse settings. However, most reasoning-oriented LMMs, including Vision-R1 (Huang et al., 2025) and R1-Onevision (Yang et al., 2025), still predominantly focus on single-image tasks and lack dedicated mechanisms for multi-image reasoning, leaving challenges with image-based multiple-choice answers largely underexplored.

Multimodal Understanding Benchmarks. Recently, numerous benchmarks have been introduced to evaluate the understanding and reasoning capabilities of LMMs. However, most focus on single-image tasks. Although several multi-image benchmarks, such as Blink (Fu et al., 2024), MUIR (Wang et al., 2024a), and MMIU (Meng et al., 2024), have emerged, they primarily assess basic perceptual abilities like caption recognition and object counting, which require limited reasoning. As LMMs advance, these benchmarks fall short in measuring deeper reasoning capabilities. In contrast, VisioMath introduces a more challenging multi-image mathematical reasoning benchmark featuring option-containing images, enabling a more comprehensive assessment of models' multi-image reasoning abilities.

Mathematical Reasoning Benchmarks. Various datasets have been proposed to evaluate the mathematical capabilities. Text-based benchmarks such as GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) are widely used. To evaluate mathematical reasoning requiring visual understanding, such as geometry and function graph analysis, several multimodal datasets have recently emerged, for example, Math-Verse (Zhang et al., 2024c), Math-Vista (Lu et al., 2024), and Math-Vision (Wang et al., 2024b). Nonetheless, as LMMs advance in multi-image reasoning, these single-image-focused benchmarks are increasingly inadequate for evaluating their full capabilities. In response, recent research efforts have begun to explore more complex multi-image reasoning scenarios that better reflect the real-world demands of mathematical problem-solving. Despite recent advances, a key limitation persists: existing multi-image benchmarks such as MathVerse-mv (Li et al., 2024b) and MV-Math (Wang et al., 2025) often neglect figure-based answer options, which are common in mathematics domain (e.g., geometry problems with diagrammatic options). This gap underscores the need for new benchmarks that support figure-based multi-image reasoning.

5 Conclusion and Limitation

We introduce VisioMath, a benchmark designed to evaluate multimodal reasoning in contexts where answer options consist of multiple, highly similar diagrams. This benchmark fills a critical gap in existing evaluation frameworks, which rarely consider the challenges of comparative reasoning across visually confusable candidates. Our experiments reveal that current LMMs perform poorly under these conditions: accuracy declines sharply with increasing inter-image similarity, and frequent errors stem from multi-image–text misalignment. Controlled shuffling experiments further show that many models rely on positional heuristics, exposing fundamental weaknesses in their reasoning mechanisms. We further explore alignment-oriented data augmentation and multi-image CoT finetuning. Results demonstrate that these strategies yield substantial gains, even under limited data regimes, indicating that relatively lightweight interventions can enhance LMMs' capacity for robust visual–textual binding.

While VisioMath provides a rigorous evaluation of multi-image, diagram-based reasoning in mathematics, our current benchmark is limited to K–12 math topics. Extending this benchmark to other domains, such as physics, engineering diagrams, or chemistry molecular structures, would test LMMs' ability to generalize multi-image reasoning across diverse visual-semantic contexts.

6 Broader Impact

VisioMath highlights critical limitations in current LMMs, particularly in fine-grained visual–text alignment and figure-based visual reasoning. By providing a targeted evaluation platform, it can guide the development of more accurate multimodal models, benefiting educational applications, intelligent tutoring systems, and diagram understanding in STEM disciplines. However, as with any benchmark, there is a risk of overfitting models to its specific structures; care must be taken to ensure that improvements reflect genuine reasoning capabilities rather than dataset-specific heuristics. Overall, we envision VisioMath supporting both model development and pedagogical research, fostering AI systems that can more effectively interpret and reason over complex visual information.

ETHICS STATEMENT

This research is based on a dataset compiled from publicly available papers from Chinese high school and college entrance examinations administered between 2002 and 2023. All data were sourced and processed in compliance with applicable laws and institutional regulations. During the curation process, we implemented a systematic filtering protocol to identify and remove any potentially harmful, offensive, or otherwise inappropriate content. Consequently, the final dataset used in this work is considered ethically sound and suitable for academic research.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we provide comprehensive documentation of our methodology. Section 2.1 details the construction and annotation of our dataset. Appendix A documents the experimental settings, model version, concrete implementations of Strategy 1 and Strategy 2, and all hyperparameters used for fine-tuning and evaluation. The construction of our CoT reasoning data is described in Appendix B, and the full set of prompts for both generation and evaluation is available in Appendix C. Together, these materials enable independent researchers to replicate our evaluation and reproduce the reported results.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, and et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

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

ByteDance. Seed1.6 tech introduction, 2024. URL https://research.doubao.com/en/seed1_6.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions, 2023. URL https://arxiv.org/abs/2311.12793.

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

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. URL https://arxiv.org/abs/2412.05271.

Zijian Chen, Wei Sun, Yuan Tian, Jun Jia, Zicheng Zhang, Jiarui Wang, Ru Huang, Xiongkuo Min, Guangtao Zhai, and Wenjun Zhang. Gaia: Rethinking action quality assessment for ai-generated videos, 2024b. URL https://arxiv.org/abs/2406.06087.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, and Ice Pasupat. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Google DeepMind. Gemini 2.0 flash, 2025a. URL https://deepmind.google/technologies/gemini/flash/.

Google DeepMind. Gemini 2.0 flash-thinking, 2025b. URL https://deepmind.google/technologies/gemini/flash-thinking/.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, and Angela Fan. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive, 2024. URL https://arxiv.org/abs/2404.12390.

Team GLM. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024. URL https://arxiv.org/abs/2406.12793.

Conghui He, Zhenjiang Jin, Chao Xu, Jiantao Qiu, Bin Wang, Wei Li, Hang Yan, Jiaqi Wang, and Dahua Lin. Wanjuan: A comprehensive multimodal dataset for advancing english and chinese large models, 2023. URL https://arxiv.org/abs/2308.10755.

Zheqi He, Xinya Wu, Pengfei Zhou, Richeng Xuan, Guang Liu, Xi Yang, Qiannan Zhu, and Hua Huang. Cmmu: A benchmark for chinese multi-modal multi-type question understanding and reasoning. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 830–838. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/92. URL https://doi.org/10.24963/ijcai.2024/92. Main Track.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021. URL https://arxiv.org/abs/2103.03874.

Jiaxing Huang and Jingyi Zhang. A survey on evaluation of multimodal large language models, 2024. URL https://arxiv.org/abs/2408.15769.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models, 2025. URL https://arxiv.org/abs/2503.06749.

Zhipeng Huang, Zhizheng Zhang, Zheng-Jun Zha, Yan Lu, and Baining Guo. Relationvlm: Making large vision-language models understand visual relations, 2024. URL https://arxiv.org/abs/2403.12801.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning, 2024. URL https://arxiv.org/abs/2405.01483.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, March 2020. ISSN 1573-1405. doi: 10.1007/s11263-020-01316-z. URL http://dx.doi.org/10.1007/s11263-020-01316-z.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions, 2024. URL https://arxiv.org/abs/2408.12637.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024a. URL https://arxiv.org/abs/2408.03326.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024b. URL https://arxiv.org/abs/2407.07895.

Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, Ying Tai, Wankou Yang, Yabiao Wang, and Chengjie Wang. A survey on benchmarks of multimodal large language models, 2024c. URL https://arxiv.org/abs/2408.08632.

Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, Guanzhou Chen, Yinan He, Zhangwei Gao, Erfei Cui, Jiashuo Yu, Hao Tian, Jiasheng Zhou, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Zhenxiang Li, Pei Chu, Yi Wang, Min Dou, Changyao Tian, Xizhou Zhu, Lewei Lu, Yushi Chen, Junjun He, Zhongying Tu, Tong Lu, Yali Wang, Limin Wang, Dahua Lin, Yu Qiao, Botian Shi, Conghui He, and Jifeng Dai. Omnicorpus: A unified multimodal corpus of 10 billion-level images interleaved with text, 2024d. URL https://arxiv.org/abs/2406.08418.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

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

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2024. URL https://arxiv.org/abs/2310.02255.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models, 2024. URL https://arxiv.org/abs/2408.02718.

Meta. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation, 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

OpenAI. Chatgpt. https://chat.openai.com/, 2023. Large language model.

OpenAI. Hello gpt-4o, 2024a. URL https://openai.com/index/hello-gpt-4o/.

OpenAI. Gpt-4 technical report, 2024b. URL https://arxiv.org/abs/2303.08774.

OpenAI. Introducing gpt-4.1 in the api, 2025. URL https://openai.com/index/gpt-4-1/.

Jaewoo Park, Jungyang Park, Dongju Jang, Jiwan Chung, Byungwoo Yoo, Jaewoo Shin, Seonjoon Park, Taehyeong Kim, and Youngjae Yu. Explain with visual keypoints like a real mentor! a benchmark for multimodal solution explanation, 2025. URL https://arxiv.org/abs/2504.03197.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Yifan Zhang, Xiao Zong, Yida Xu, Muxi Diao, Zhimin Bao, Chen Li, and Honggang Zhang. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024. URL https://arxiv.org/abs/2407.01284.

Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. From pixels to prose: A large dataset of dense image captions, 2024. URL https://arxiv.org/abs/2406.10328.

Kai Sun, Yushi Bai, Ji Qi, Lei Hou, and Juanzi Li. Mm-math: Advancing multimodal math evaluation with process evaluation and fine-grained classification, 2024. URL https://arxiv.org/abs/2404.05091.

Yanpeng Sun, Shan Zhang, Wei Tang, Aotian Chen, Piotr Koniusz, Kai Zou, Yuan Xue, and Anton van den Hengel. Mathglance: Multimodal large language models do not know where to look in mathematical diagrams, 2025. URL https://arxiv.org/abs/2503.20745.

Doubao Team. Doubao-1.5-pro, January 2025a. URL https://team.doubao.com/zh/special/doubao_1_5_pro.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024a. URL https://arxiv.org/abs/2403.05530.

Gemma Team. Gemma 3 technical report, 2025b. URL https://arxiv.org/abs/2503.19786.

Qwen Team. Qvq: To see the world with wisdom, December 2024b. URL https://qwenlm.github.io/blog/qvq-72b-preview/.

Qwen Team. Introducing qwen-vl, January 2025c. URL https://qwenlm.github.io/blog/qwen-vl/.

V Team, Wenyi Hong, Wenmeng Yu, and Xiaotao Gu. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. Muirbench: A comprehensive benchmark for robust multi-image understanding, 2024a. URL https://arxiv.org/abs/2406.09411.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*, 2024b. URL https://arxiv.org/abs/2402.14804.

Peijie Wang, Zhong-Zhi Li, Fei Yin, Xin Yang, Dekang Ran, and Cheng-Lin Liu. Mv-math: Evaluating multimodal math reasoning in multi-visual contexts, 2025. URL https://arxiv.org/abs/2502.20808.

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

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2024d. URL https://arxiv.org/abs/2311.03079.

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

Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. Logical closed loop: Uncovering object hallucinations in large vision-language models, 2024a. URL https://arxiv.org/abs/2402.11622.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024b. URL https://arxiv.org/abs/2412.10302.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts, 2024. URL https://arxiv.org/abs/2407.04973.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization, 2025. URL https://arxiv.org/abs/2503.10615.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, Los Alamitos, CA, USA, June 2024. IEEE Computer Society. doi: 10.1109/CVPR52733.2024.00913. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.00913.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Hang Yan, Conghui He, Xingcheng Zhang, Kai Chen, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output, 2024a. URL https://arxiv.org/abs/2407.03320.

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

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*, 2024c.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift:a scalable lightweight infrastructure for fine-tuning, 2024. URL https://arxiv.org/abs/2408.05517.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text, 2023. URL https://arxiv.org/abs/2304.06939.

Yi Zong and Xipeng Qiu. GAOKAO-MM: A Chinese human-level benchmark for multimodal models evaluation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8817–8825, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.521. URL https://aclanthology.org/2024.findings-acl.521.

A Experiment Details

A.1 Experiment Setting

All experiments were conducted on a Linux server equipped with two NVIDIA H800 GPUs (each with 80GB of memory). The Python version used in the experiments was 3.9.20, while the version of vllm library was 0.8.1, respectively. Each model evaluation was performed in a zero-shot setting with deterministic decoding (temperature=0). Due to frequent updates and improvements, closed-source models often undergo version changes that can significantly impact evaluation results. Even subtle updates may alter model behavior, performance, or prompt adherence. As such, the results reported in this benchmark are tied to the specific versions used during our evaluation. To ensure transparency and reproducibility, Table 5 lists the exact release dates or version identifiers of all closed-source models evaluated in VisioMath. Readers should be aware that performance discrepancies may arise when using newer or older versions of the same models.

Specifically, we employ the same prompt template across all models to eliminate prompt-induced variance, and fix the decoding temperature to 0 to promote deterministic outputs. Accuracy serves as the primary evaluation metric, measuring the proportion of correctly answered instances. We utilize GLM4-Flash (GLM, 2024) to extract the options from the responses generated by LMMs. In scenarios where the model fails to produce a valid answer, i.e., none of the standard options (A, B, C, or D) can be reliably identified, its response is classified as invalid. Such cases are treated as incorrect predictions in the final accuracy computation.

To ensure a fair comparison, we adopted consistent prompting strategies across the three input types: Original, Strategy 1, and Strategy 2. For Strategy 1, we horizontally concatenated all images with zero-padding. For Strategy 2, we extend each option image by adding a 50-pixel-high strip at the bottom, matching the width of the image, and insert the corresponding option letter (A, B, C, or D) within the strip. Examples of these configurations are illustrated in Figure 5. In Strategy 3, we adopted the Supervised Fine-Tuning (SFT) training strategy on QwenVL2.5-3B. Using a single H800 GPU and the ms-swift framework (Zhao et al., 2024), we set the batch size to 2, the learning rate to 1e-4, and the gradient accumulation steps to 4. The training was conducted over 336 steps.

Table 5: Version information or release dates of evaluated closed-source models.

Model	**Version (release date)**
GPT-4o (OpenAI, 2024a)	2024-11-20
GPT-4.1 (OpenAI, 2025)	2025-04-14
Gemini2-flash (DeepMind, 2025a)	2024-12-11
Gemini2-flash-thinking (DeepMind, 2025b)	2025-01-21
Gemini 2.5 Pro (Comanici et al., 2025)	2025-06-17
QwenVL-max (Team, 2025c)	2025-04-08
QwenVL-plus (Team, 2025c)	2025-01-25
Doubao-1.5-Vision-pro (Team, 2025a)	2025-03-28
Seed1.6 (ByteDance, 2024)	2025-08-15
GLM4V-plus (GLM, 2024)	2025-01-11
GLM4.5V (Team et al., 2025)	2025-08-11

A.2 Full Experimental Results

Due to space limitations in the main text, here we report the full evaluation results of various LMMs on the VisioMath benchmark in Table 6 and Table 7.

B Multi-Image CoT Fine-Tuning

This section explains how CoT reasoning data was constructed, including description generation by QwenVL-Max, refinement by DeepSeek3.1, and filtering strategies. We construct a specialized multi-image chain-of-thought (CoT) dataset through a structured three-stage pipeline to enhance model performance.

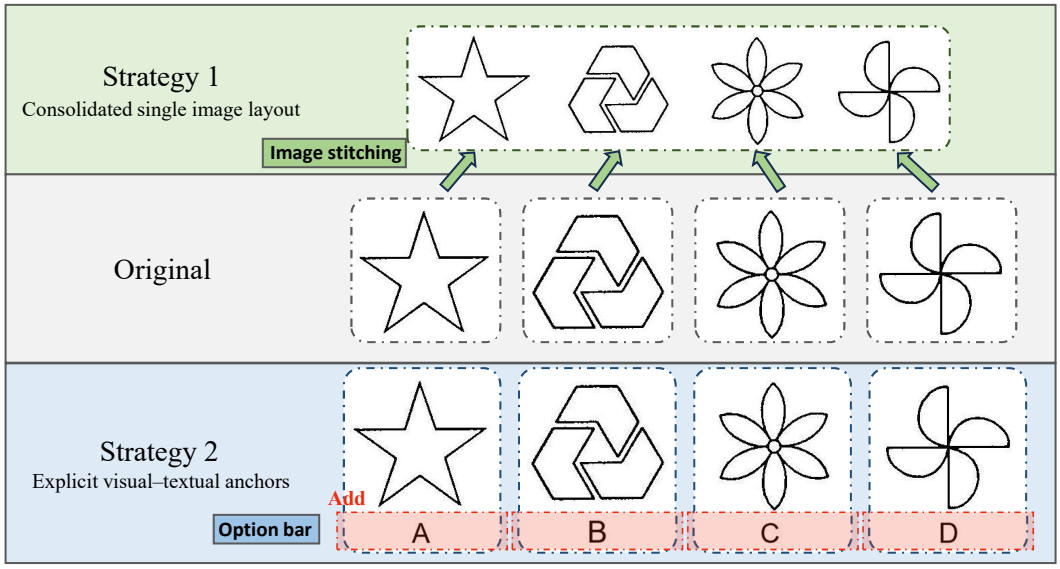


Figure 5: Illustrating the image format of original, strategy 1, and strategy 2 in our experiments.

Stage 1 (*Problem collection*). We crawled 1,072 mathematical single-choice questions that contain more than four pictures from the internet to serve as the raw problem pool.

Stage 2 (*CoT Sample Generation*). Initial reasoning paths and descriptive captions are produced for each problem using QwenVL-Max with an Image-Caption–Style Prompt. These outputs, together with the original questions, are then fed into DeepSeek V3.1 via a CoT Data Generation Prompt to generate refined reasoning trajectories and corresponding answers. Samples are subsequently filtered based on answer correctness, resulting in 500 high-quality multi-image CoT examples.

Stage 3 (*Dataset Expansion*). To increase both the scale and diversity of the dataset, a Option Shuffling Prompt strategy is applied, expanding the dataset from 0.5k to 1.3k samples.

This three-stage pipeline ensures that the final dataset contains both high-quality reasoning examples and sufficient data scale, providing a robust foundation for effective model training.

C Prompt Templates

We employ the same prompt template across all models to eliminate prompt-induced variance. Specifically, we use five types of system prompts in our paper:

- **Original Answer Prompt**: The baseline system instruction that is uniformly appended to all models prior to evaluation, serving to standardize response format and output scope.
- **Option Shuffling Prompt**: A variant of the Option Shuffling Prompt in which the correspondence between options and images is completely deranged, designed to test and mitigate the model's reliance on positional priors, and used for synthetic data generation.
- **Answer Extraction Prompt**: A prompt used to guide the LLM in extracting and normalizing the final answer from the model's output (e.g., mapping free-form text or reasoning steps to discrete options such as A/B/C/D).
- **Image-Caption–Style Prompt**: A prompt that instructs the MLLM to generate concise, comparable textual descriptions and preliminary analyses for each image, serving as a cross-modal representation bridge.
- **CoT Data Generation Prompt**: A prompt that integrates the question, image captions, and MLLM-provided reasoning trajectories to produce high-quality chain-of-thought rationales and final answers, which can be leveraged for data augmentation and fine-tuning.

The detailed prompt texts are shown in Table 8.

D Error Analysis

This section presents a detailed analysis of errors, categorizing them into four types, reporting their distributions, and providing representative examples.

Image-Text Misalignment (36%). These errors occur when GLM-4.5V fails to correctly capture the semantic correspondence between textual options and visual content. For example, in Figure 6, the model misinterprets the relationship between the image and the answer options, incorrectly treating the reference image as Option A.

Vision Recognition Error (34%). Vision recognition errors reflect the model's difficulty in accurately perceiving visual information. As shown in Figure 7, GLM-4.5V fails to correctly interpret the shapes of the unfolded cubes corresponding to Options B and C.

Reasoning Error (20%). Reasoning errors arise when GLM-4.5V does not correctly follow logical steps or underlying problem constraints. For instance, in Figure 7, the model incorrectly assumes that the depicted line graph necessarily satisfies the definition of a function.

Knowledge Error (10%). Knowledge errors occur when GLM-4.5V lacks relevant domain knowledge or produces outdated/inaccurate information. For example, in Figure 6, the model erroneously interprets the top view of a sphere as a circle with a visible center point.

E Use of LLMs

In this work, large language models (LLMs) were utilized as supporting tools to aid in the revision and polishing of certain text segments during manuscript preparation. All model-generated content was thoroughly examined, revised, and refined by the authors to guarantee accuracy and compliance with academic writing standards. Notably, the study's conceptual framework, methodological design, data analysis, and interpretation of findings were conducted entirely by the authors, without dependence on automated systems. The authors bear full responsibility for the originality, validity, and integrity of the research presented in this paper.

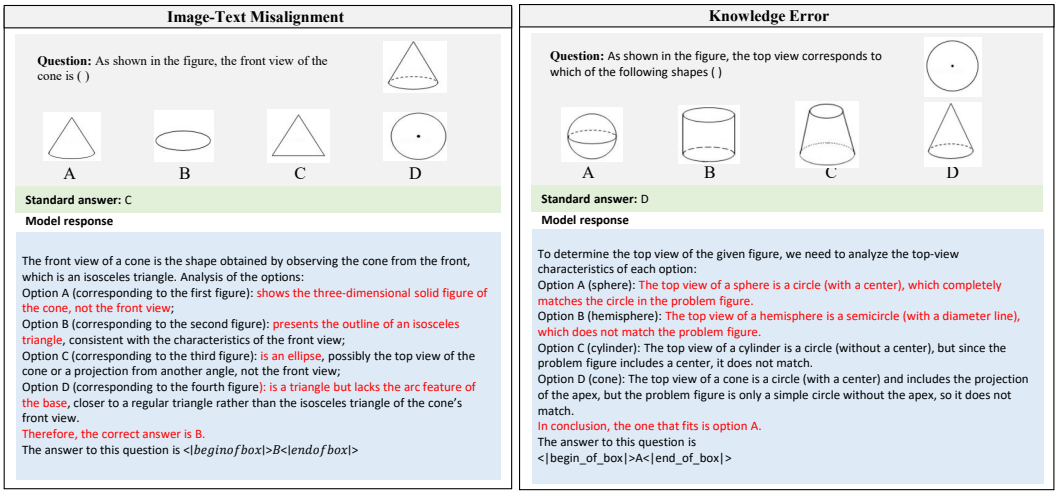


Figure 6: Example error cases for image-text misalignment and knowledge error.

Reasoning Error

Question: Among the following curves, which represents y as a function of x ( )

A B C D

Standard answer: D

Model response

According to the definition of a function, for every value of x in the domain, there must be a uniquely determined y corresponding to it. To judge whether a curve represents a function, one can use the "vertical line test": if any line perpendicular to the x-axis intersects the curve at most once, then the curve represents y as a function of x.
Option A: In the case of a straight line, each x corresponds to a unique y, satisfying the definition of a function.
Option B: In the case of a circle, there exist certain values of x (e.g., x=1) that correspond to two different y values (y=1 and y=-1), which does not satisfy the definition of a function.
Option C: In the case of a hyperbola, similarly, there exist x values corresponding to multiple y values (e.g., x=1 corresponds to y=1 and y=-1), which does not satisfy the definition of a function.
Option D: Although the curve shows an overall increasing trend, it is necessary to check whether there are intervals where one x corresponds to multiple y values. If the curve "turns back" in some interval (i.e., the same x corresponds to multiple y values), it would not satisfy the definition of a function. Based on common problem types, such curves are usually not considered to represent functions (for example, a cubic function may have local regions with multiple values).
In conclusion, only option A satisfies the definition of a function.
The answer to this question is <|begin_of_box|>A<|end_of_box|>

Vision Recognition Error

Question: Which of the following figures cannot be folded into a cube ( )

A B C D

Standard answer: B

Model response

To determine which figure cannot be folded into a cube, rely on the features of cube nets:
Option A: Belongs to the "1-4-1" type; it is a valid cube net.
Option B: Also a "1-4-1" type (four faces in the middle with one above and one below), which can be folded into a cube.
Option C: Contains a "field" grid structure (four small squares forming a larger square), which is forbidden for cube nets and cannot be folded into a cube.
Option D: Belongs to the "1-3-2" type and can be folded into a cube.
Therefore, the one that cannot be folded into a cube is option C.
The answer to this question is <|begin_of_box|>C<|end_of_box|>

Figure 7: Example error cases for reasoning error and vision recognition error.

Table 6: Performance comparison on VisioMath with results categorized based on GT position.

Models \GT position	Avg	Question stem w/o images					Question stem with images				
		Avg	A	B	C	D	Avg	A	B	C	D
Random	25.6	25.4	24.0	25.6	23.0	28.6	26.0	22.8	27.6	28.4	25.6
Closed LMMs											
GLM4V-plus (GLM, 2024)	27.9	30.2	28.1	33.5	31.6	27.5	25.4	39.4	22.9	26.0	11.1
QwenVL-plus (Bai et al., 2023)	32.9	39.1	27.0	59.9	43.4	25.5	26.3	7.5	26.2	34.8	40.1
QwenVL-max (Bai et al., 2023)	44.1	53.4	35.2	62.6	62.5	50.2	34.1	31.1	34.1	32.8	38.6
GPT-4o (OpenAI, 2024a)	45.9	54.7	55.6	56.4	54.7	52.5	36.5	47.3	30.4	36.3	30.4
GPT4.1 (OpenAI, 2025)	52.6	61.6	72.4	59.9	60.2	56.1	42.8	54.8	39.3	43.7	31.9
Gemini2-flash-thinking (DeepMind, 2025b)	53.2	61.2	80.6	59.9	58.6	50.3	44.6	57.3	43.0	43.1	32.9
GLM-4.5V (Team et al., 2025)	53.7	69.1	71.9	75.8	68.4	61.2	37.2	46.5	42.5	31.4	26.6
Gemini2-flash (DeepMind, 2025a)	55.5	65.1	78.1	59.9	65.2	59.6	45.2	57.7	34.5	38.7	47.8
Doubao-1.5-Vision-pro (Team, 2025a)	66.3	75.6	78.6	78.0	76.6	70.2	56.4	73.4	55.6	48.0	45.4
Seed1.6-Thinking (ByteDance, 2024)	72.3	85.7	**90.3**	**87.2**	82.4	83.9	58.0	71.8	53.7	44.6	59.4
Gemini 2.5 Pro (Comanici et al., 2025)	**80.9**	**86.2**	89.2	84.6	**85.2**	**86.3**	**75.2**	**78.8**	**77.6**	**75.0**	**68.6**
Multi-Image input LMMs											
deepseekvl2-tiny (Wu et al., 2024b)	23.5	21.6	45.9	29.1	15.2	2.7	25.6	58.5	27.1	6.9	4.3
InternVL2.5-2B-MPO (Chen et al., 2024a)	23.9	24.9	15.3	26.9	33.2	22.4	22.9	13.7	28.5	36.3	14.5
InternVL2.5-2B (Chen et al., 2024a)	24.6	27.1	12.8	25.5	36.3	30.2	21.9	10.3	26.2	38.2	15.0
Llama3.2-11B-Vison (Dubey et al., 2024)	25.3	26.2	30.6	24.7	39.8	10.6	24.2	30.2	23.4	36.3	6.3
Idefics3-8B-llama (Laurençon et al., 2024)	25.4	26.1	20.9	55.9	19.1	10.6	24.6	39.8	32.7	11.3	11.6
Qwen2.5-VL-3B-instruct (Bai et al., 2025)	25.4	26.1	51.0	40.5	14.5	5.9	24.7	18.3	**70.1**	5.4	4.3
Phi3.5-vision (Abdin et al., 2024)	25.7	25.3	73.5	22.0	14.1	2.4	26.2	**78.8**	13.6	3.4	0.5
deepseekvl2-small (Wu et al., 2024b)	26.6	32.0	42.8	28.6	30.4	28.2	20.8	38.6	12.6	12.3	16.9
Mantis-8B-Idefics2 (Jiang et al., 2024)	27.9	30.8	24.0	17.6	42.1	36.5	24.8	22.0	7.0	32.4	39.1
InternVL2.5-4B (Chen et al., 2024a)	28.2	30.4	23.5	37.0	34.0	26.3	25.8	22.0	36.0	34.8	10.6
InternVL2.5-4B-MPO (Chen et al., 2024a)	28.4	30.9	12.8	33.0	31.6	42.4	25.6	9.5	27.1	47.1	21.7
MiniCPM-o-2.6 (Abdin et al., 2024)	29.3	34.6	40.8	36.1	31.3	31.8	23.6	36.5	23.8	15.7	15.9
R1-Onevison-7B (Yang et al., 2025)	29.6	35.0	38.8	37.4	34.8	30.2	23.7	22.0	32.2	28.9	11.6
MiniCPM-V-2.6 (Abdin et al., 2024)	29.7	33.0	31.6	30.4	30.9	38.4	26.1	30.0	26.2	16.7	30.9
InternVL2.5-8B (Chen et al., 2024a)	29.9	33.1	26.4	31.7	50.4	22.0	26.6	30.7	16.4	40.7	18.4
InternVL2.5-8B-MPO (Chen et al., 2024a)	30.9	35.9	25.0	48.5	39.5	29.4	25.5	23.7	25.2	34.3	19.3
QvQ-72B-Preview (Team, 2024b)	30.9	36.2	45.4	36.1	44.0	31.4	25.3	31.5	28.0	26.0	14.5
Qwen2-VL-72B-instruct (Wang et al., 2024c)	31.7	38.2	15.8	29.5	78.1	23.1	24.5	2.5	9.8	83.8	6.8
Qwen2.5-VL-7B-instruct (Bai et al., 2025)	32.7	39.5	30.1	58.1	39.8	29.8	25.3	8.7	28.5	32.4	34.3
Gemma3-27B (Team, 2025b)	35.3	43.7	67.9	40.1	33.6	38.4	26.2	40.2	24.8	12.3	25.1
Vision-R1-7B (Huang et al., 2025)	36.7	43.7	47.4	57.3	38.7	33.7	29.2	24.5	52.3	29.4	10.6
Qwen2.5-VL-32B-instruct (Bai et al., 2025)	41.8	51.2	**68.3**	53.2	47.7	39.6	31.8	65.1	22.9	16.2	17.4
Qwen2.5-VL-72B-instruct (Bai et al., 2025)	43.7	53.5	36.2	63.9	61.3	49.8	33.0	29.9	37.8	29.9	35.2
Llama4-Maverick-17B-128E-FP8 (Meta, 2025)	**66.9**	**70.1**	64.8	**71.8**	**71.1**	**71.8**	**63.4**	61.4	61.7	**77.0**	**54.1**
Single-Image input MLMs											
LLaVA-v1.6-vicuna-7B (Liu et al., 2024a)	20.7	22.6	21.4	8.4	**32.4**	26.3	18.7	29.5	2.8	23.0	18.4
MiniCPM-v2.5 (Abdin et al., 2024)	21.0	21.7	28.1	13.2	12.1	34.1	20.2	28.2	15.4	6.4	**29.5**
LLaVA-onevision-7B (Li et al., 2024a)	22.7	19.8	**79.1**	7.0	3.9	1.6	26.0	70.1	20.6	3.4	2.4
LLaVA-v1.6-mistral-7B (Liu et al., 2024a)	23.0	19.9	73.5	3.1	12.9	0.8	**26.3**	**78.4**	2.3	16.7	0.0
LLaVA-v1.5-7B (Liu et al., 2023)	23.7	23.6	23.5	19.4	12.5	**38.4**	23.8	33.6	17.3	14.2	28.5
GLM4V-9B (GLM, 2024)	23.9	**25.6**	19.4	**31.7**	31.6	18.8	22.2	10.3	**33.2**	**26.0**	20.8
LLaVA-v1.6-vicuna-13B (Liu et al., 2024a)	**24.4**	23.0	50.5	2.2	5.1	**38.4**	26.0	66.4	0.0	2.9	28.5

Table 7: Performance comparison on VisioMath with results categorized based on image similarity.

Models \Image similarity	Avg	[0.16,0.68]	(0.68,0.87]	(0.87,0.96]	(0.96,1]
Random	25.6	23.6	24.4	27.8	27.1
Closed LMMs					
GLM4V-plus (GLM, 2024)	27.9	29.6	32.9	23.3	25.8
QwenVL-plus (Bai et al., 2023)	32.9	33.3	37.8	32.4	28.2
QwenVL-max (Bai et al., 2023)	44.1	47.3	50.2	41.3	37.6
GPT-4o (OpenAI, 2024a)	45.9	53.8	50.9	40.0	39.1
GPT-4.1 (OpenAI, 2025)	52.6	65.8	56.4	42.9	45.1
Gemini2-flash-thinking (DeepMind, 2025b)	53.2	63.6	58.9	48.2	42.2
GLM-4.5V (Team et al., 2025)	53.7	68.7	59.3	44.2	44.7
Gemini2-flash (DeepMind, 2025a)	55.5	66.7	59.8	49.3	46.2
Doubao-1.5-Vision-pro (Team, 2025a)	66.3	74.9	68.2	60.2	62.0
Seed1.6-Thinking (ByteDance, 2024)	72.3	82.4	74.2	66.2	66.4
Gemini 2.5 Pro (Comanici et al., 2025)	**80.9**	**86.2**	**83.8**	**76.7**	**76.9**
Multi-Image input LMMs					
DeepSeekVL2-tiny (Wu et al., 2024b)	23.5	23.3	24.0	24.4	22.4
InternVL2.5-2B-MPO (Chen et al., 2024a)	23.9	24.0	27.6	24.0	20.2
InternVL2.5-2B (Chen et al., 2024a)	24.6	24.2	28.9	22.7	22.7
Llama3.2-11B-Vison (Dubey et al., 2024)	25.3	23.3	27.8	26.4	23.6
Idefics3-8B-llama (Laurençon et al., 2024)	25.4	26.9	26.0	22.7	26.0
Qwen2.5-VL-3B-instruct (Bai et al., 2025)	25.4	26.7	27.6	24.4	22.9
Phi3.5-vision (Abdin et al., 2024)	25.7	23.6	28.7	27.8	22.9
DeepSeekVL2-small (Wu et al., 2024b)	26.6	30.7	29.6	24.9	21.3
Mantis-8B-Idefics2 (Jiang et al., 2024)	27.9	32.2	28.9	24.7	26.0
InternVL2.5-4B (Chen et al., 2024a)	28.2	28.9	31.8	27.3	24.7
InternVL2.5-4B-MPO (Chen et al., 2024a)	28.4	28.2	34.0	26.2	25.1
MiniCPM-o-2.6 (Abdin et al., 2024)	29.3	34.9	35.3	24.4	22.4
R1-Onevison-7B (Yang et al., 2025)	29.6s	21.9	32.2	28.9	11.6
MiniCPM-V-2.6 (Abdin et al., 2024)	29.7	30.7	34.9	28.4	24.7
InternVL2.5-8B (Chen et al., 2024a)	29.9	32.4	31.8	29.6	26.0
InternVL2.5-8B-MPO (Chen et al., 2024a)	30.9	35.6	37.1	25.8	25.1
QvQ-72B-Preview (Team, 2024b)	30.9	37.3	38.0	25.3	23.1
Qwen2-VL-72B-instruct (Wang et al., 2024c)	31.7	35.5	37.8	26.0	27.1
Qwen2.5-VL-7B-instruct (Bai et al., 2025)	32.7	33.6	37.8	29.8	29.6
Gemma3-27B (Team, 2025b)	35.3	43.3	41.2	29.6	26.4
Vision-R1-7B (Huang et al., 2025)	36.7	46.7	38.9	30.4	30.9
Qwen2.5-VL-32B-instruct (Bai et al., 2025)	41.8	50.0	46.2	38.4	32.7
Qwen2.5-VL-72B-instruct (Bai et al., 2025)	43.7	47.1	50.8	38.0	38.7
Llama4-Maverick-17B-128E-Instruct-FP8 (Meta, 2025)	**66.9**	**63.6**	**70.0**	**65.8**	**68.2**
Single-Image input LMMs					
LLaVA-v1.6-vicuna-7B (Liu et al., 2024a)	20.7	22.2	24.4	17.8	18.4
MiniCPM-V-2.5 (Abdin et al., 2024)	21.0	21.7	21.3	20.6	20.2
LLaVA-onevision-7B (Li et al., 2024a)	22.7	22.2	22.4	25.6	20.9
LLaVA-v1.6-mistral-7B (Liu et al., 2024a)	23.0	21.8	**26.0**	23.6	20.5
LLaVA-v1.5-7B (Liu et al., 2023)	23.7	23.3	25.3	24.9	21.1
GLM4V-9B (GLM, 2024)	23.9	**26.7**	23.5	23.3	**22.0**
LLaVA-v1.6-vicuna-13B (Liu et al., 2024a)	**24.4**	24.0	**26.0**	**26.0**	21.8

Table 8: This table presents the prompts used for process evaluation and answer generation by various LMMs in the VisioMath benchmark.

Phase	Input	Prompt
Answer Extraction (GLM4-Flash)	Model's response	You are an AI assistant that helps me extract the answers to single-choice questions. You will be provided with an answer. Your task is to find the final option of the model. If the model's answer is meaningless, output Z. You should output a single uppercase letter, such as A, B, C, D (if they are valid options), or Z. **Example 1**: Answer: According to the question description and all related pictures, option A is the correct answer. Option A is a centrally symmetric figure because its four vertices are all symmetric, while the vertices of options B, C, and D are not symmetric. Output: A **Example 2**: Answer: A. Sphere B. Circle C. Disc D. Circle Output: Z **Example 3**: **Answer**: {model answer} **Output**:
Answer Generation (LMMs)	Question Diagrams	Please solve a single-choice math question. The last four pictures are respectively the pictures for options A, B, C, and D. Select the correct answer from A, B, C, and D based on the question description and all relevant pictures.
Option Shuffling Generation (LMMs)	Question Diagrams	Please solve a single-choice math question. The last four pictures are respectively the pictures for options B, C, Dand A. Select the correct answer from A, B, C, and D based on the question description and all relevant pictures.
Image Caption Generation (LMMs)	Question Diagrams	I have multiple images and a question that I want you to answer. I need you to strictly follow the format with three specific sections: SUMMARY, CAPTION and REASONING. To explain further: In SUMMARY, briefly explain what steps you'll take to solve the problem. In CAPTION, describe the contents of all the images, wrapping each image description inside tags like <image1></image1>, <image2></image2>, etc. In REASONING, outline a step-by-step thought process you would use to solve the problem based on the images. <SUMMARY> [Summarize how you will approach the problem...] </SUMMARY> <CAPTION> <image1>... </image1> <image2>... </image2> </CAPTION> <REASONING> [Provide a chain-of-thought, logical explanation of the problem. This should outline step-by-step reasoning based on all the images.] </REASONING>
CoT Data Generation (DeepSeek-V3.1)	Question Caption	Please solve this multiple-choice math question and answer in English. The last four images correspond to options A, B, C, and D respectively. Based on the question description and all relevant images, select the correct answer from A, B, C, and D.