Supplementary Material - MATHVERSE: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems?

Renrui Zhang*‡[1,2], Dongzhi Jiang*[1], Yichi Zhang*[2], Haokun Lin[2], Ziyu Guo[3]
Pengshuo Qiu[2], Aojun Zhou[1], Pan Lu[4], Kai-Wei Chang[4]
Yu Qiao[2], Peng Gao[2], and Hongsheng Li[1,5]

[1] CUHK MMLab
[2] Shanghai AI Laboratory
[3] CUHK MiuLar Lab
[4] University of California, Los Angeles
[5] CPII under InnoHK
{renruizhang, dzjiang, ziyuguo}@link.cuhk.edu.hk
gaopeng@pjlab.org.cn hsli@ee.cuhk.edu.hk

Overview

A Related Work

Multi-modal Large Language Models (MLLMs), building upon the prevalence of Large Language Models (LLMs) [4, 27, 44, 52, 53], have become increasingly prominent for incorporating multi-modalities [28,48,64,66,67]. They extend LLMs to tackle a diverse range of tasks and domains, including the mainstream 2D images [1,14,29,32] and other modalities, such as 3D point clouds [22,26,56], audio [23,51], video [6,63], and robotics [33,37,38,57]. Noteworthy examples like OpenAI's GPT-4V [46] and Google's Gemini [20] exhibit exceptional visual understanding and reasoning capabilities, setting new benchmarks in multi-modal performance. However, their closed-source nature poses a barrier to the broader application and development of MLLMs. Concurrently, another line of work is

*Equal contribution ‡Project lead †Corresponding author

dedicated to exploring advanced MLLMs open-source to the community. Prior efforts like LLaMA-Adapter [18, 65], LLaVA [31, 35, 36], and MiniGPT-4 [11, 70] leverage a frozen CLIP [48] model for image encoding, and inject the visual cues into LLaMA [52] for multi-modal instruction tuning. The subsequent mPLUG-Owl [58, 59], Qwen-VL [3], InternLM-XComposer [15], and SPHINX [19, 34] further push the frontier of MLLMs in understanding and generalizing across visual contexts. MAVIS [68] introduces the first mathematical visual instruction-tuning paradigm with two newly curated datasets, MAVIS-Caption and MAVIS-Instruct, demonstrating improved reasoning skills. Despite comprehensive benchmarks [16, 21, 30, 39, 55, 60] on general visual instruction-following scenarios, the specific potential of MLLMs for visual mathematical problem-solving remains under-explored. In this paper, we introduce the MATHVERSE benchmark to comprehensively evaluate the visual mathematical reasoning skills of MLLMs, providing unique perspectives for future research directions.

Mathematical Reasoning Benchmarks have emerged as a significant area of focus, posing considerable challenges for large foundational models, e.g., LLMs and MLLMs. Initially, datasets in this realm are designed to address basic algebraic [25] and arithmetic [49] word problems, which are relatively limited in scope and volume. Subsequent efforts, including MATH [25], GSM8K [13], and MMLU [24], expand the range and quality of textual mathematical problems. These datasets feature a broader spectrum of difficulties, establishing a robust benchmark for the evaluation of general and math-specific LLMs [17, 43, 54, 62, 69]. Besides the text-only assessment, there is a growing demand for comparable, high-quality benchmarks for evaluating mathematical problem-solving in visual contexts, with the rapid progress of MLLMs. There are prior attempts, such as GeoQA [9], UniGeo [7], and Geometry3K [41], which focused exclusively on geometric problems. The recently proposed MathVista [40] broadens the scope to incorporate a variety of multi-modal tasks involving mathematical reasoning, and MMMU [61] covers college-level questions demanding intricate, domain-specific knowledge. However, our analysis identifies three main shortcomings within the current visual math benchmarks, as elaborated in Section 1 of the main paper. Therefore, we propose MATHVERSE specialized in the multi-modal mathematical evaluation of MLLMs, comprising twelve subjects, six problem versions, and 20K test samples. Our objective is to thoroughly investigate whether and how much MLLMs genuinely interpret visual diagrams for mathematical reasoning.

B Additional Experimental Details

More Implementation Details. We conduct all experiments on NVIDIA A100 GPUs. For different MLLMs, we select their latest models and best-performing configurations for evaluation. Table 3 presents the source of the models used in MATHVERSE. As the text-only LLMs can only take text questions as input, we evaluate them with the first three problem versions, i.e., Text Dominant, Text Lite, and Text Only. For the 'w/o' results, we utilize the template in

Table 1: Input Prompt of MLLMs for Response Generation. We adopt two different prompts for the free-form and multiple-choice questions. Note that these prompts are used for five problem versions except for the Vision-only version.

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*}

Table 2: Input Prompt for Vision-only Problems. Especially for the Vision-only version without textual input, we add *"According to the question shown in the image"* at the beginning of the prompt, and remove the *"Question:"* at the end.

Question	Prompt
Free-form Question	According to the question shown in the image, please first conduct reasoning, and then answer the question and provide the final value, e.g., 1, 2.5, 300, at the end.
Multiple-choice Question	According to the question shown in the image, please first conduct reasoning, and then answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.

MathVista [40] to prompt GPT-4 [45] for answer extraction, and directly score the final answer without the intermediate reasoning process.

Prompt for Response Generation. We adopt two types of prompts respectively for the free-form and multiple-choice questions, as shown in Table 1. We inspire the Chain-of-Thought (CoT) reasoning capabilities of MLLMs by using the phrase *"first conduct reasoning"*. Especially for the Vision-only problem version in Table 2, we add *"According to the question shown in the image"* at the beginning to remind MLLMs to read the questions rendered within diagrams, where the textual input for MLLMs only contains the prompt itself.

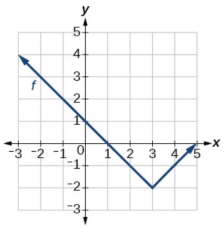
Prompt for the CoT Evaluation. Our proposed CoT evaluation contains two steps, i.e., key-step extraction and multi-step scoring, which prompt GPT-4 [45] and GPT-4V [46], respectively. The input configuration is listed in Table 4. We utilize the text-only GPT-4 in the first step to extract multiple key steps within the model's unstructured output, without feeding the question information. In

Table 3: The Source of the Models Used in MathVerse.

Model	Release Time	Source
ChatGPT [47]	2022-11	https://platform.openai.com/
GPT-4 [45]	2023-03	https://platform.openai.com/
Qwen-VL-Plus [3]	2023-11	https://help.aliyun.com/zh/dashscope/developer-reference/vl-plus-quick-start
Gemini-Pro [20]	2023-12	https://ai.google.dev/
Qwen-VL-Max [3]	2024-01	https://help.aliyun.com/zh/dashscope/developer-reference/vl-plus-quick-start
GPT-4V [46]	2023-09	https://platform.openai.com/
LLaMA-Adapter V2 [18]	2023-04	https://github.com/OpenGVLab/LLaMA-Adapter/tree/main/llama_adapter_v2_multimodal7b
LLaVA-1.5 [35]	2023-10	https://huggingface.co/liuhaotian/llava-v1.5-13b
MiniGPT-v2 [11]	2023-10	https://github.com/Vision-CAIR/MiniGPT-4
mPLUG-Owl2 [59]	2023-11	https://huggingface.co/MAGAer13/mplug-owl2-llama2-7b
G-LLaVA [17]	2023-12	https://github.com/pipilurj/G-LLaVA/tree/main
ImageBind-LLM [23]	2023-05	https://github.com/OpenGVLab/LLaMA-Adapter/tree/main/imagebind_LLM
ShareGPT4V [12]	2023-11	https://huggingface.co/Lin-Chen/ShareGPT4V-13B
SPHINX-Plus [34]	2023-11	https://huggingface.co/Alpha-VLLM/LLaMA2-Accessory/tree/main/finetune/mm/SPHINX/SPHINX-v2-1k
LLaVA-NeXT [36]	2024-01	https://huggingface.co/liuhaotian/llava-v1.6-vicuna-13b
SPHINX-MoE [19]	2024-01	https://huggingface.co/Alpha-VLLM/LLaMA2-Accessory/tree/main/finetune/mm/SPHINX/SPHINX-MoE
InternLM-XComposer2 [15]	2024-01	https://huggingface.co/internlm/internlm-xcomposer2-vl-7b

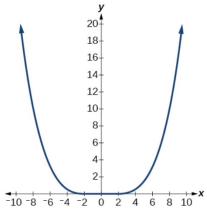
Annotation:

A piecewise function f:
when $x <= 3,\ f(x) = -x + 1$,
when $x > 3,\ f(x) = x - 5$.



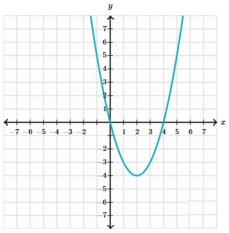
Annotation:

A U-shape curve,
passes the origin (0, 0),
opens upward.



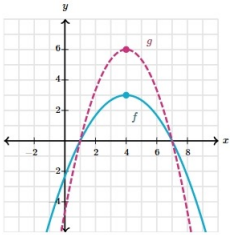
Annotation:

A quadratic function $y = (x-2)^2 - 4$,
goes through (0, 0), (4, 0),
opens upward with a vertex (2, 4).



Annotation:

Two quadratic functions going through (1, 0) and (7, 0),
$f(x) = -(1/3)(x-1)(x-7)$ in blue with a vertex is (4, 3),
$g(x) = -(2/3)(x-1)(x-7)$ in dashed red with a vertex is (4, 6).



Annotation:

A linear function $y = 0.5x + 2$,
passes (-4, 0), (0, 2) and (2, 3),
the shaded region is above the line.

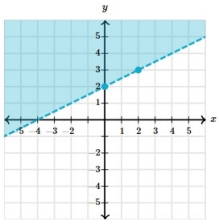


Fig. 1: Manual Annotations for Function Problems in MathVerse. We provide detailed annotations, e.g., function expression and properties, for the diagrams of 534 function problems, which benefits the accuracy of GPT-4V [46] for CoT evaluation.

the second step, we input the extracted key-step reasoning and all the available content related to the problem into GPT-4V, allowing for a holistic assessment, including diagram interpretation, logical reasoning, and numerical computation. In Figure 1, we showcase the manual annotation for critical information within functional diagrams, e.g., function expression and properties. This assists GPT-4V in evaluating the visual perception accuracy of MLLMs for function graphs.

Human Performance Assessment. We recruit ten qualified college students specifically for the evaluation of human performance on MathVerse. These individuals are kept separate from the data curation stage, eliminating the possibility of them encountering the solutions beforehand. We allocate to each student the questions from a specific problem version. This strategy is to prevent them from gaining additional information from another version to answer questions, e.g., leveraging the textual *Implicit Property* from the Text-lite version to solve Text-intensive problems. They are asked to directly provide the final answer

Table 4: Configuration for the CoT Evaluation Strategy. We conduct two evaluation phases, respectively by prompting the text-only GPT-4 [45] and GPT-4V [46]. The symbol 'XXX' denotes the given one-shot sample, abbreviated for brevity. The 'Annotation' in the second phase is only required for function problems.

Phase	Input	Prompt
Key-step Extraction (GPT-4)	Model Output	I will give you a detailed solving procedure or a single answer for a math problem. If it is a procedure, you need to extract the key solution steps and list them accordingly in markdown syntax. If it is just a single answer, output the answer directly. Here are examples: – Model output: XXX – Extracted: 1. XXX 2. XXX 3. XXX – Model output: 2.2 – Extracted: The single answer is 2.2 Here is what you need to extract: – **Model output:** {*model output*} – **Extracted:**
Multi-step Scoring (GPT-4V)	Extracted Steps Question Diagram Answer Annotation	I will first give you a visual math problem, including the question, diagram, ground-truth answer, and detailed annotation of the diagram, and then give you a model output containing multiple key solution steps. Please think step by step and output the Average score, along with the Final answer score in the end, as described below: – Average score: Evaluate, based on the given question, answer, diagram, and diagram annotation, whether each solution step is correct in logical reasoning, visual perception, and numerical computation, with an incorrect score of 0 and a correct score of 1. Then, calculate the average score of multiple steps. – Final answer score: Match the model's final answer with the ground truth answer, scoring 1 if it matches and 0 if it doesn't. – If the model output only includes a single step or answer, the Average score and Final answer score are the same. – **Question:** {*question*} – **Ground-truth answer:** {*answer*} – **Diagram annotation:** {*annotation*} – **Model output:** {*extracted steps*} – **Average score:** – **Final answer score:**

without detailed reasoning. Therefore, we do not report the CoT evaluation results for human performance, alongside the 'Random Chance' baseline.

C Additional Experimental Analysis

Closed-source MLLMs are Better-performed. From the performance in both accuracy and CoT evaluation, we observe a consistently better performance achieved by closed-source MLLMs than open-source ones. Despite the gap with humans, GPT-4V attains the leading position among MLLMs, showcasing superior mathematical capabilities over problem versions and subjects. MAVIS-7B is the best-performing open-source MLLM, surpassing Gemini-Pro and Qwen-VL-Max, while still lagging behind GPT-4V, suggesting the improvement space.

LLMs Achieve Competitive Results to MLLMs. Utilizing solely question texts as input, two LLMs, i.e., GPT-4 and ChatGPT, attain superior accuracy to most MLLMs in Text Dominant and Lite versions. Even in the absence of redundant *Descriptive Information* within Text-lite problems, GPT-4 outperforms LLaVA-NeXT and MAVIS-7B models. These findings not only indicate the strong mathematical reasoning skills of LLMs, but further emphasize the deficiencies in diagram interpretation of existing MLLMs. Importantly, the performance of GPT-4 is only exceeded by GPT-4V, which demonstrates that a satisfactory diagram perception capability can enhance the problem-solving for visual mathematics.

Mathematical Training Benefits the Performance. In addition to foundational visual instruction-following datasets, both InternLM-XComposer2 and LLaVA-NeXT extend their training regimes to include specialized mathematical problems that are either text-only or visual, such as MathQA [2], Geometry3K [41], and MathInstruct [62]. This approach of math-specific tuning contributes to their leading performance in MathVerse. Furthermore, G-LLaVA fine-tunes LLaVA-1.5 by a large-scale visual geometric dataset containing 170K enriched problems. This targeted refinement can improve several fields ('Len', 'Angle', and 'Apply') within the plane geometry subject. However, since G-LLaVA's fine-tuning data does not include problems of analytic geometry, solid geometry, and functions, it harms the related results of LLaVA-1.5 due to catastrophic forgetting. Moreover, due to a three-stage training strategy and large-scale multi-modal mathematical data, MAVIS-7B exhibits superior performance over other MLLMs. This phenomenon underscores the critical role of developing extensive, high-quality visual math datasets for the effective training of MLLMs.

Discrepancy Between CoT and Accuracy Evaluation Scores. As illustrated by the two tables in the main paper, the CoT scores for MLLMs, in most cases, are much higher than accuracy scores. This observation demonstrates that our proposed CoT evaluation strategy identifies numerous correct intermediate reasoning steps, despite the final incorrect answer, highlighting the effectiveness

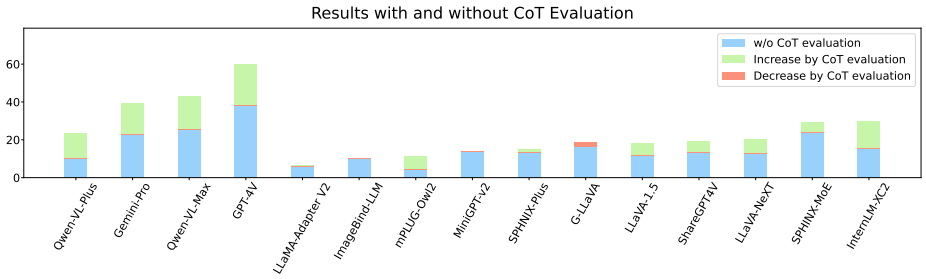


Fig. 2: Results with and without CoT Evaluation in MathVerse. Referring to Table **??**, we denote the 'w/o' results in blue pillars, and highlight the increase and decrease magnitude with 'CoT-E' by green and red, respectively.

of fine-grained assessment. In Figure 2, we present the statistics of variance between the CoT and accuracy scores within different MLLMs. Although GPT-4V attains top-tier performance, it exhibits a pronounced gap concerning the evaluation of CoT reasoning quality. Conversely, SPHINX-MoE showcases favorable precision among open-source MLLMs, while preserving a relatively lower variance of two evaluation methods compared to InternLM-XComposer. This indicates its consistent step-by-step reasoning throughout the solving process.

D More Dataset Details

D.1 Data Curation

This paper engages twelve expert annotators for data curation, consisting of senior undergraduate and graduate students from across the globe with a strong background in science. In collaboration with the authors, they are required to mainly complete five tasks concerning data collection, categorization, quality review, problem version transformation, and function diagram annotation.

Data Collection. We comprehensively collect visual math problems from existing datasets [10,41,50] and public question repositories[1,2,3]. We specifically select high-quality plane geometric problems from current benchmarks, which showcase various question types, moderate question length, diverse diagram styles, and appropriate solving difficulty. For the manually collected problems of three subjects (plane geometry, solid geometry, and functions), we apply the Mathpix tool[4] to accurately extract the question texts, diagrams, explanations, and answers from the website. We strictly comply with copyright and licensing rules,

[1]https://homework.study.com
[2]https://www.ixl.com/math
[3]https://mathspace.co/us
[4]https://mathpix.com

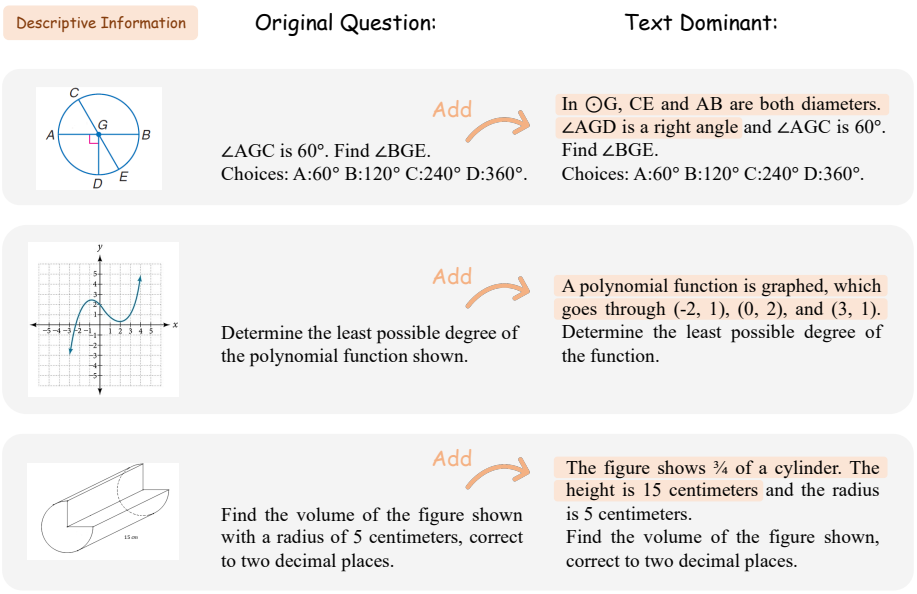


Fig. 3: Manual Annotations for *Descriptive Information* in MathVerse. For some collected problems, we are required to supplement additional *Descriptive Information* (highlighted in red) to distinguish the Text-dominant version.

ensuring that we refrain from using data from sites that forbid copying and redistribution. After the initial collection, we obtain around 3.5K visual math problems, with 1.5K from existing datasets and 2K newly collected.

Data Categorization and Review. We first ask the human annotators to categorize the problems into three primary subjects, i.e., plane geometry, solid geometry, and functions. Within each subject, according to the definitions in Section D.2, the math problems are further divided into twelve fine-grained categories. At the same time, we meticulously review the collected dataset. We manually rectify the problems with incorrect answers and discard the problems with multiple diagrams, visual solutions, and too much similar content to others. Finally, 2,612 high-quality math problems with paired diagrams are preserved for MathVerse, spanning diverse subjects and subfields.

Transformation of Problem Versions. Given the three types of textual information within questions, human annotators rigorously transform each problem into six different versions as discussed in Section 2.2 of the main paper. We utilize Microsoft PowerPoint to annotate the diagrams in the Vision-dominant version, and employ Matplotlib to render the questions onto the diagrams in the Vision-only version. As illustrated in Figure 3, for problems with minimal *Descriptive Information*, we manually enhance the question text with additional contextual description about the diagram to differentiate the Text-dominant ver-

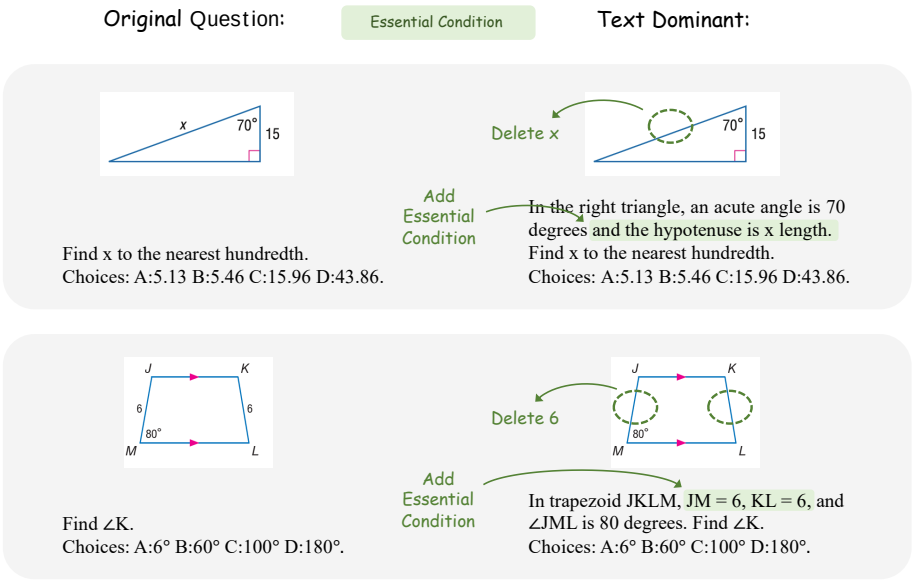


Fig. 4: Manual Annotations for *Essential Condition* in MathVerse. For the original problems shown, we transfer some of the *Essential Condition* from diagrams to question texts (highlighted in green) to mark the Vision-dominant version.

sion. In the case of questions in Figure 4, where the *Essential Condition* has been fully depicted in the diagrams, we remove some of this content from the diagram and incorporate it into the text to mark the Vision-dominant version.

D.2 Subject and Subfield Definition

The visual math problems within MathVerse encompass three primary subjects, plane geometry, solid geometry, and functions, alongside twelve finer-grained subfields, which comprehensively evaluate the diagram understanding and mathematical reasoning capabilities of MLLMs.

Plane Geometry is a fundamental area that explores the properties and relations of points, lines, and surfaces in a two-dimensional plane. This subject delves into concepts such as angles, triangles, circles, and polygons, offering a rich context for assessing the spatial comprehension and logical deduction skills of MLLMs. We divide it into five subfields, as exemplified in Figure 5:

– **Length** focuses on the measurement and comparison of distances between points. This subfield includes understanding the properties of lines, segments, and their use in determining the perimeters of geometric shapes, which is foundational for MLLMs to solve plane geometry problems.

Plane Geometry:

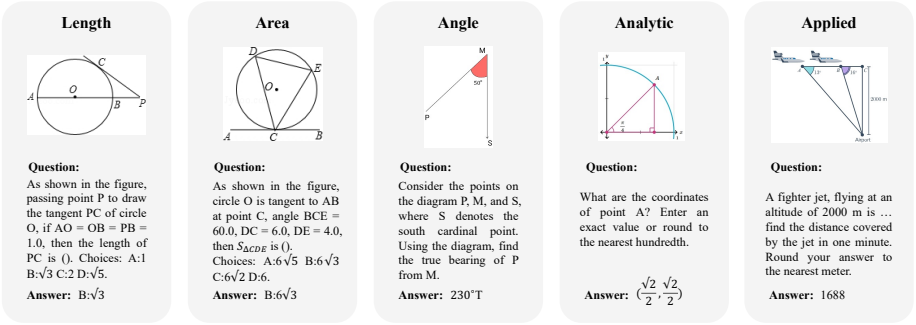


Fig. 5: Examples of Five Subfields in Plane Geometry, spanning Length, Area, Angle, Analytic, and Applied Geometry problems. We showcase the Text-lite version.

- **Area** examines the size of two-dimensional surfaces. It encompasses calculating the areas of various shapes, such as triangles, rectangles, circles, and more complex polygons, by applying specific formulas and principles, which is crucial for comprehending the concept of space within geometry.
- **Angle** involves the study of angles and their properties, including different types of angles (acute, right, and obtuse), angle measurement, and the relationships between angles, particularly in polygons. This subfield demands the advanced spatial perception capacity of MLLMs.
- **Analytic Geometry**, also known as coordinate geometry, merges algebra and geometry to solve geometric problems using coordinate systems, exploring the calculation and reasoning of equations for geometric shapes. MLLMs are evaluated on their coordinate identification and algebraic capabilities.
- **Applied Geometry** relate to the application of geometric principles to solve real-world and theoretical problems. It challenges MLLMs to first understand the background information within questions, and apply their knowledge of lengths, areas, angles, and analytic geometry for problem-solving.

Solid Geometry focuses on the study of three-dimensional objects that have depth, length, and width, thereby offering a more complex and enriched exploration of spatial structures. This subject investigates a variety of shapes such as cubes, cylinders, spheres, and pyramids, and assesses MLLMs to tackle questions concerning the volume, surface area, and geometric properties of these solids. This subject contains three subfields, as exemplified in Figure 6:

- **Length**, extending from the 2D counterpart, focuses on measuring the edges and curves that define three-dimensional objects. It involves determining the linear distance between points in space, the perimeters of bases of solids, and the height or depth of objects. This measurement is a foundational element for MLLMs in analyzing geometric solids.

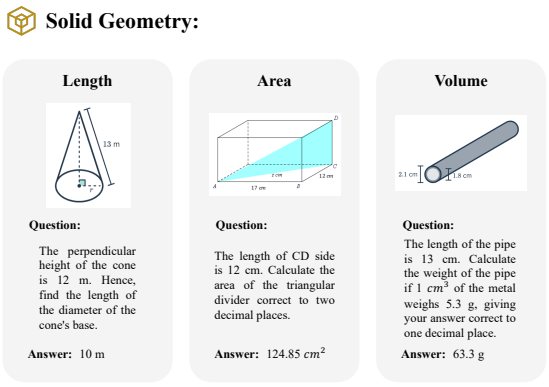


Fig. 6: Examples of Three Subfields in Solid Geometry, spanning Length, Area, and Volume problems. We showcase the Text-lite version.

- **Area** encompasses the calculation of the total area covered by the outer surfaces of solids. This normally requires MLLMs to break down complex shapes into several simpler components for area calculation in plane geometry, assessing their spatial and logical reasoning performance.
- **Volume** pertains to measuring the space enclosed within three-dimensional objects. This demands MLLMs to precisely identify the geometric solids and apply accurate formulas to calculate the volume, which evaluates their mathematical knowledge application and calculation skills.

Functions involve analyzing mathematical functions to understand the relationship between variables. These challenges range from simple tasks, like calculating a function value for a given input, to more complex scenarios, such as exploring the behavior and representation of various function types. We evaluate MLLMs by four types of function problems, exemplified in Figure 7:

- **Function Coordinate** focuses on interpreting and extracting coordinate-level information from graphical representations of functions. It includes tasks such as identifying specific coordinate values of points on the graph and observing intersection points between functions and axes, which test the MLLM's basic proficiency in functional visual perception.
- **Function Property** emphasizes the model's capacity to discern and deduce the inherent properties of functions from their graphs, such as symmetry, asymptotes, extrema (maximum and minimum points), and intervals of increase or decrease. These problems can reveal the understanding of MLLMs for the deeper characteristics of functions.
- **Function Expression** refers to the direct analysis using the algebraic expressions of functions, widely including linear, quadratic, polynomial, exponential, logarithmic, and piece-wise functions. It challenges MLLMs to extract specific function expressions and apply transformations, bridging the gap between abstract mathematical reasoning and visual interpretation.

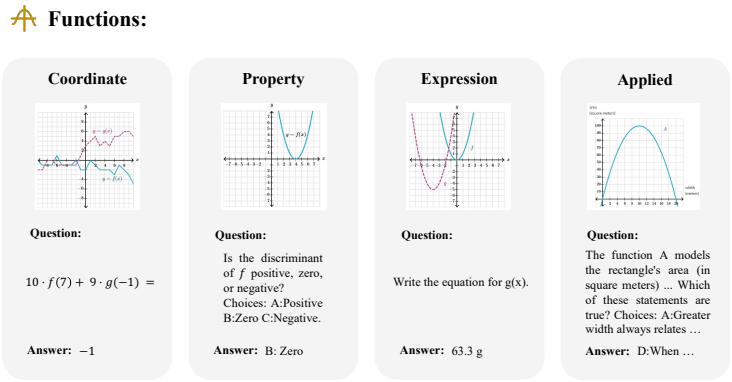


Fig. 7: Examples of Four Subfields in Functions, spanning Function Coordinate, Property, Expression, and Applied problems. We showcase the Text-lite version.

– **Applied Function**, similar to the applied geometry problems, requires MLLMs to leverage their functional knowledge and theorems in real-world scenarios, e.g., modeling economic data, predicting physical phenomena, and calculating probabilities. This assesses the MLLM's capabilities to understand functions in both theoretical contexts and practical situations.

D.3 Detailed Statistics of MathVerse

More Data Statistics. In Table 5, we provide a more detailed data statistics of MathVerse. Therein, the 534 newly annotated questions refer to all the function problems, for which we meticulously annotate critical functional information, as depicted in Figure 1. The number of newly annotated diagrams represents the 5,224 math problems in the Vision-dominant and Vision-only versions. For these problems, we respectively integrate the *Essential Condition* and all textual content with the diagrams. We also list the numbers of multiple-choice answers, where A, B, C, and D are almost uniformly distributed.

Problem Length Variance. In Table 6, we highlight the variance in question and answer lengths across the five problem versions in MathVerse, excluding the Vision-only category due to its absence of text. For both word and character levels, as we remove the pre-defined textual elements (*Descriptive Information, Implicit Property*, and *Essential Condition*), the maximum and average lengths of questions decrease accordingly, while the answer lengths remain the same. In Figure 8, we visualize the word-level variation of question length for the three problem versions: Text Dominant, Text Lite, and Vision Dominant. By progressively omitting *Descriptive Information* and *Essential Condition* from the Text-dominant version, we observe a clear downward trajectory for the question length distribution and average values.

Table 5: Statistics of MathVerse.

Statistic	Number
Total questions	2,612
- Subjects/subfields	3/12
- Multiple-choice questions	1,631 (62.4%)
- Free-form questions	981 (37.6%)
- **Newly collected questions**	**1,236 (47.3%)**
- Existing-dataset questions	1,376 (52.7%)
- **Questions with explanations**	**1,236 (47.3%)**
- **Newly annotated questions**	**534 (20.4%)**
Multiple-choice question	
- Proportion of answer A	585 (22.4%)
- Proportion of answer B	828 (31.7%)
- Proportion of answer C	703 (26.9%)
- Proportion of answer D	444 (17.0%)
- Proportion of answer E&F	52 (2.0%)
Total test samples	**15,672**
- **Newly annotated samples**	**10,448 (66.7%)**
- **Newly annotated diagrams**	**5,224 (33.3%)**
- Samples of each version	2,612 (16.7%)
Number of unique images	2,420 (92.6%)
Number of unique questions	2,573 (98.5%)
Number of unique answers	847 (32.4%)

Table 6: Length of Different Problem Versions in MathVerse.

Problem Version	Word	Character
Text Dominant & Text Only		
- Maximum question length	203	1,311
- Maximum answer length	17	102
- Average question length	35.7	204.8
- Average answer length	1.4	6.3
Text Lite		
- Maximum question length	179	1,173
- Maximum answer length	17	102
- Average question length	22	133.8
- Average answer length	1.4	6.3
Vision Intensive		
- Maximum question length	171	1,126
- Maximum answer length	17	102
- Average question length	18.8	116.8
- Average answer length	1.4	6.3
Vision Dominant		
- Maximum question length	176	1,132
- Maximum answer length	17	102
- Average question length	17.6	123.5
- Average answer length	1.4	6.3

E Comparison to Current Benchmarks

In this section, we offer a detailed comparison between MathVerse and existing multi-modal mathematical benchmarks, i.e., geometry-specific benchmarks [5, 8, 10, 42, 50], MathVista [40], and MMMU [61], from the following four aspects:

The Investigation of Diagram Interpretation Capacity. As discussed in Figure 1 of the main paper, the math problems in most existing datasets contain excessive redundant information in textual content, which is repetitive to the visual elements in diagrams. This issue enables MLLMs to potentially bypass the process of visual understanding, and thereby cannot determine whether and how much MLLMs truly interpret the math diagram. In contrast, our MathVerse includes six problem versions with different information content across text and vision. By comparing the performance variance between different problem versions, we can thoroughly investigate the mathematical diagram interpretation capabilities of MLLMs for the first time.

Evaluation Approach. Previous benchmarks adopt a simple True or False metric to score the response from MLLMs, which lacks fine-grained information and intermediate reasoning assessment, as analyzed in Figure 2 of the main paper. In contrast, MathVerse adopts a unique CoT evaluation strategy by examining each crucial solution step within the model output. This approach not only unveils the CoT reasoning quality of MLLMs, but also provides detailed error analysis, serving as valuable guidance for future enhancement.

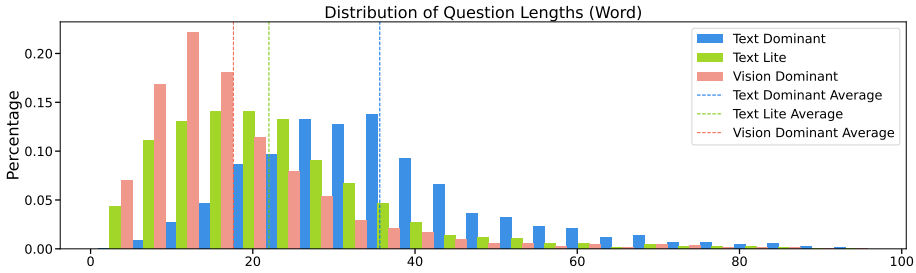


Fig. 8: Distribution of Question Length for Three Problem Versions. We exclude the *Descriptive Information* and *Essential Condition* from the Text-dominant problems, respectively creating the Text-lite and Vision-dominant versions.

The Depth and Width in Math Problems. The geometry-specific benchmarks evaluate only a limited dimension of mathematical skills in MLLMs. MathVista instead incorporates a variety of math-related question-answering tasks, e.g., textbook figures, tables, plots, charts, puzzles, and synthetic scenes, as exemplified in Figure 9. However, the integration of these peripheral tasks (covering more than 70%) might divert the focus from the specialized mathematical evaluation of MLLMs. In addition, MMMU focuses on college-level complexity, requiring advanced domain-specific knowledge, as depicted in Figure 10. Given this, the lack of profound mathematical theorems would restrict the performance of MLLMs, biasing the evaluation of logical reasoning and visual perception proficiency. Therefore, our MATHVERSE concentrates on specialized visual math problems (plane geometry, solid geometry, and functions) with a moderate difficulty (high-school level), aiming to fully exert the capabilities of MLLMs.

Total Volume of Test Samples. We summarize the size of test instances for different datasets in Table 7. As demonstrated, our MATHVERSE offers a considerably larger number of samples than others, nearly three times to MathVista and twenty times to GeoQA+, including meticulously annotated six versions of visual math problems. This contributes to a comprehensive and robust evaluation of visual mathematical reasoning capabilities.

Table 7: Number of Test Samples in Different Benchmarks.

Benchmark	GEOS	Geo3K	GeoQA+	MathVista	MMMU-Math	**MATHVERSE**
Test Samples	119	601	755	6,141	540	**15,672**

Table QA:

Items sold last week	
Item	Frequency
TV	16
tablet computer	7
speaker	32
cell phone	23
video game console	19

Making leaf rubbings	
Leaf rubbings made	Frequency
0	14
1	5
2	15
3	11
4	3
5	10
6	17

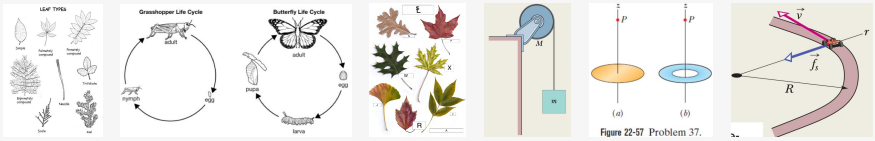
Rounds in the spelling bee	
Year	Rounds
2008	17
2009	10
2010	9
2011	15
2012	7

Athletes per country

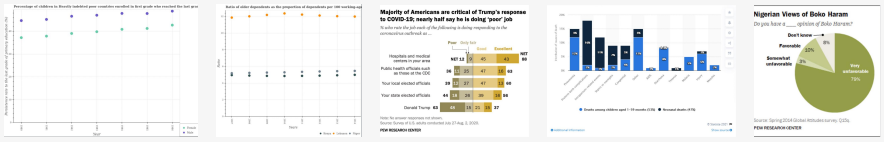
Stem	Leaf
1	1 3 3
2	0 1 2 5 7 9 9
3	4 5 7 8
4	1
5	2 4 5 5 6 6 8
6	1 6

peanut butter cookie dough	$3 per lb
double chocolate cookie dough	$3 per lb
chocolate chip cookie dough	$5 per lb
oatmeal raisin cookie dough	$7 per lb
gingersnap cookie dough	$3 per lb
snickerdoodle cookie dough	$8 per lb

Textbook and Science QA:



Plot and Chart QA:



IQ Test and Synthetic QA:

General and Icon QA:

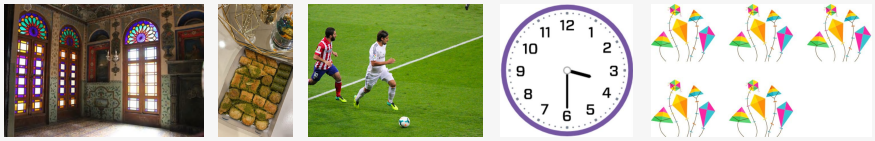


Fig. 9: Diagram Examples of Math-related Tasks in MathVista [40]. These tasks are not strongly correlated to the mathematical reasoning skills of MLLMs, probably skewing the assessment emphasis towards visual math problems.

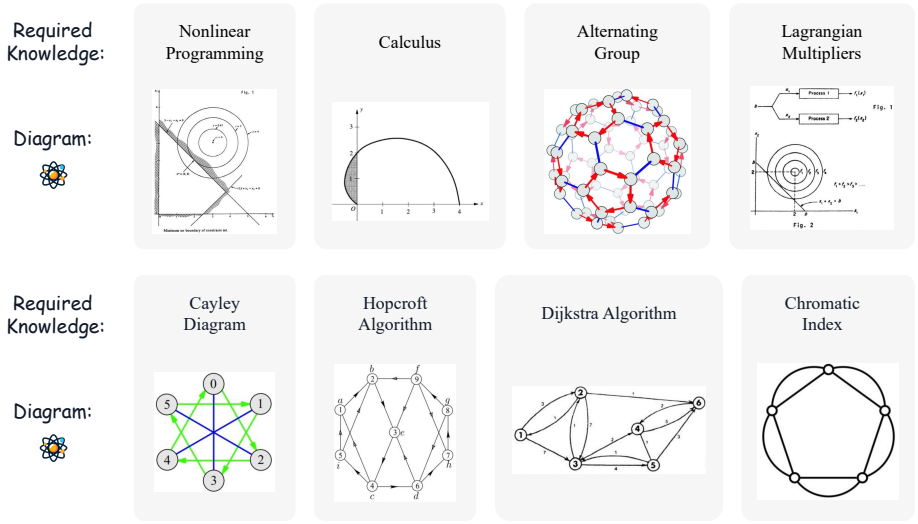


Fig. 10: Diagram Examples with Required Knowledge in MMMU [61]. These math problems demand MLLMs to comprehend college-level domain knowledge, potentially hindering them from fully exerting mathematical reasoning skills.

F Limitation and Future Work

While our MathVerse takes a step forward in the field of visual mathematical evaluation for MLLMs, it is important to recognize several limitations as follows.

We have categorized the math problems in MathVerse by various criteria, including subjects, subfields, and versions featuring differing degrees of multi-modal content. These categorization approaches evaluate the capabilities of MLLMs from multiple dimensions. Nevertheless, it is also meaningful to further divide the problems based on their difficulty levels, akin to MATH [25], a text-only benchmark defining five levels of difficulty. This additional layer of differentiation can provide deeper insights into the problem-solving abilities of MLLMs across a spectrum of challenges, which we leave as future work.

The curated dataset in MathVerse focuses on math problems in the high school level with moderate difficulty, which aims to fully demonstrate the mathematical reasoning skills within current MLLMs. However, with the advancement of architecture and training methodologies, future MLLMs have the potential to grasp more complex knowledge and theorems across a variety of domains. Therefore, there is significant value in further augmenting MathVerse with problems spanning broader complexity and disciplines, including those at the college level and within scientific fields. By transforming the expanded problems into different versions, we can facilitate a more comprehensive and robust evaluation of MLLMs for their diagram interpretation and reasoning capabilities.

Moreover, the problems in MathVerse and other current mathematical benchmarks are mainly in English. Given that some multilingual MLLMs [3, 23]

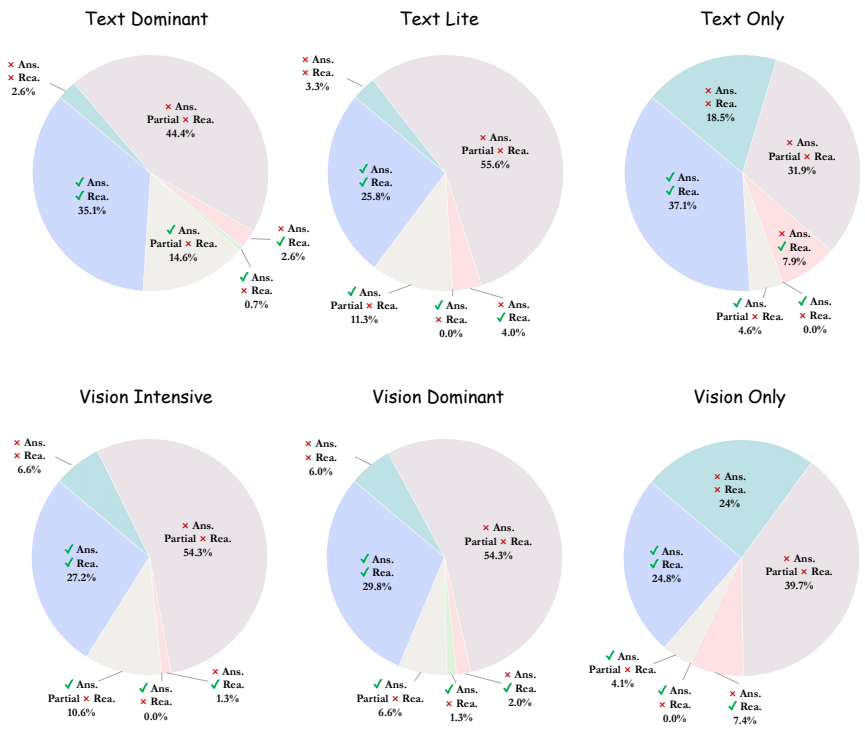


Fig. 11: Distribution of GPT-4V's [46] Errors in Reasoning and Answers. For the six problem versions in MathVerse, we provide the statistics of errors made by GPT-4V based on their occurrence in answers ('Ans.') and reasoning processes ('Rea.').

have been developed, existing evaluation cannot reveal their full capabilities when confined to a single language. The incorporation of multilingual visual math problems would not only extend the dataset's global applicability, but also enhance the assessment of MLLMs for linguistic diversity and understanding.

G Error Analysis

To delve into the fine-grained predictions, we select the best-performing MLLM, GPT-4V [46], to understand its modes of success and failure. Our proposed CoT evaluation strategy has produced a detailed assessment of model output, including step-wise scores and explanation, reducing extensive manual effort in identifying and analyzing errors. We conduct our analysis on the two-step output from the CoT evaluation across the entire dataset, focusing on two key dimensions.

Errors in Reasoning or Answer? In Figure 11, we showcase the statistics of different error distributions in six problem versions of MathVerse. We define the following six error categories: correct final answer with correct/partially correct/incorrect CoT reasoning and incorrect final answer with correct/partially

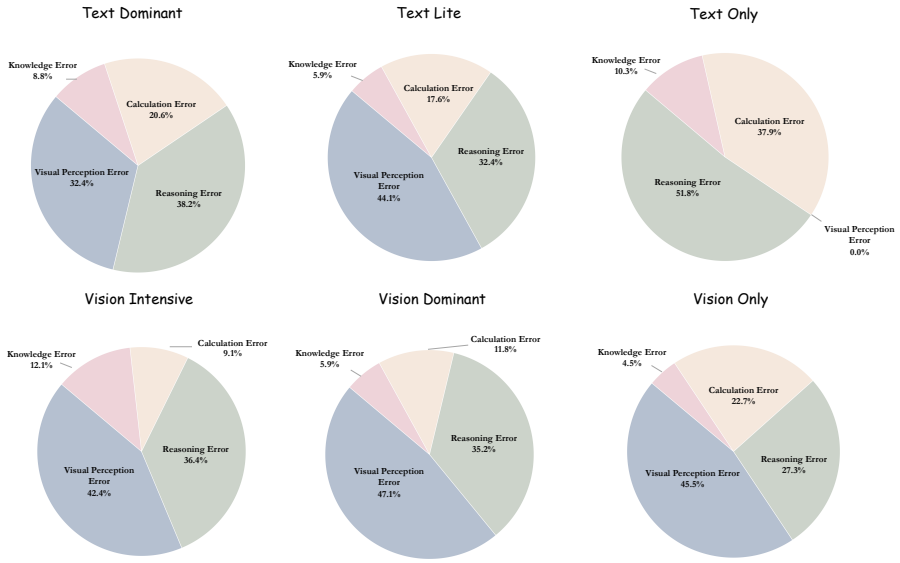


Fig. 12: Distribution of GPT-4V's [46] Errors within Different Types. We present the statistics of four error types by GPT-4V in the six problem versions, i.e., Visual Perception Error, Reasoning Error, Calculation Error, and Knowledge Error.

correct/incorrect CoT reasoning. For all six versions, the incorrect final answers are mostly caused by the partially incorrect reasoning process. In addition, a number of problems with correct answers are accompanied by partially or entirely incorrect reasoning, e.g., 15.3% in Text Dominant, which cannot be detected by the traditional True or False evaluation. As we remove the content within textual questions and enrich the visual diagram, e.g., from Text Dominant and Lite to Vision Dominant and Only, we observe a progressive increase in the error rate of 'incorrect final answer with incorrect CoT reasoning', indicating that MLLMs are challenged to conduct high-quality intermediate reasoning by capturing more information from the visual input.

What Types of Errors? To further investigate the specific error types, we survey the problems with errors that occur either within the reasoning process or the final answer. As depicted in Figure 12, we divide the errors of GPT-4V into four distinct types: visual perception error, reasoning error, knowledge error, and calculation error. Consistent with our findings in the main paper, the primary source of errors in problem-solving attributes to the inaccurate interpretation of mathematical diagrams, which significantly impedes the performance of MLLMs. For the problem versions that demand advanced diagram interpretation, e.g., Vision Dominant and Only, we observe a notable increase in the rate of visual perception errors, demonstrating an urgent need for stronger visual encoders in MLLMs. Moreover, reasoning errors also account for a considerable percentage, indicating that the logical deduction skills of MLLMs still

require improvement. As expected, knowledge errors do not significantly hinder the mathematical reasoning capabilities of MLLMs in MathVerse.

H Qualitative Examples

To ease the understanding, we offer a variety of qualitative examples in MathVerse. In Section H.1, we showcase the meticulously transformed six versions of visual math problems. In Section H.2, we compare the response of different MLLMs on Text-lite problems, including GPT-4V [46], LLaVA-NeXT [36], and SPHINX-MoE [19]. Specifically, we present the key-step extraction output by the CoT evaluation, and mark the multi-step scoring results aside. In Section H.3, we provide the response comparison of GPT-4V for three problem versions in MathVerse, i.e., Text Dominant, Text Lite, and Vision Dominant.

H.1 Comparison of Six Problem Versions

Please refer to Figures 13∼15.

H.2 Response of Different MLLMs

Please refer to Figures 16∼21.

H.3 Response of Different Problem Versions

Please refer to Figures 22∼27.

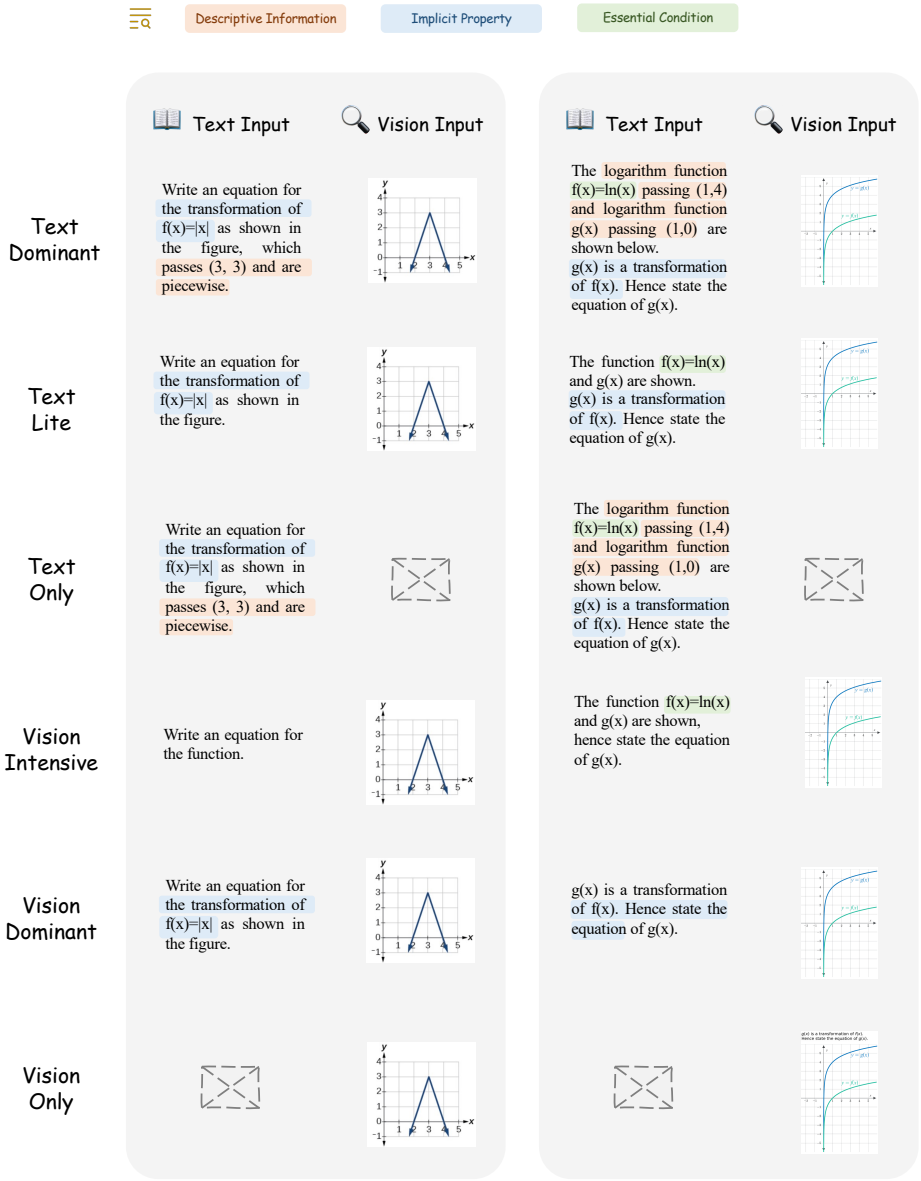


Fig. 13: Comparison of Six Problem Versions in MathVerse.

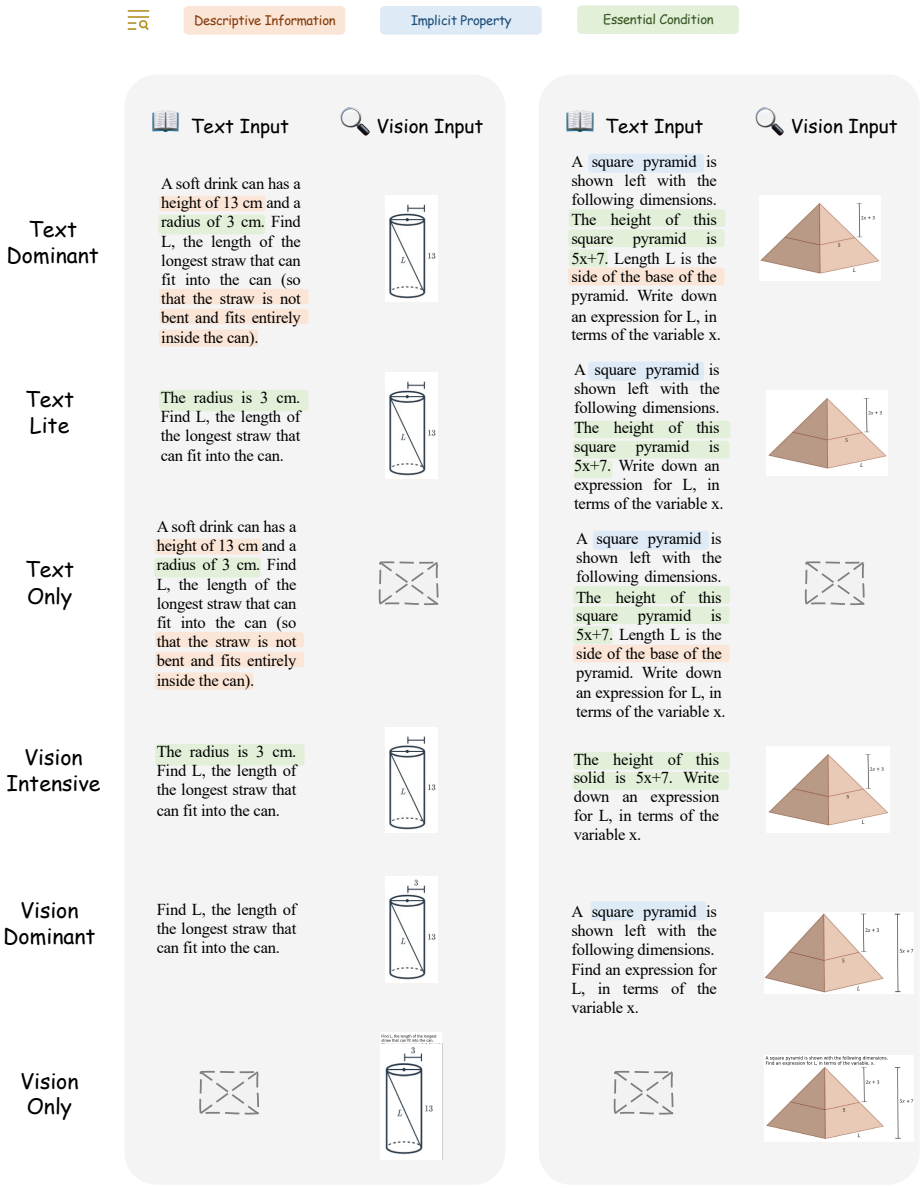


Fig. 14: Comparison of Six Problem Versions in MathVerse.

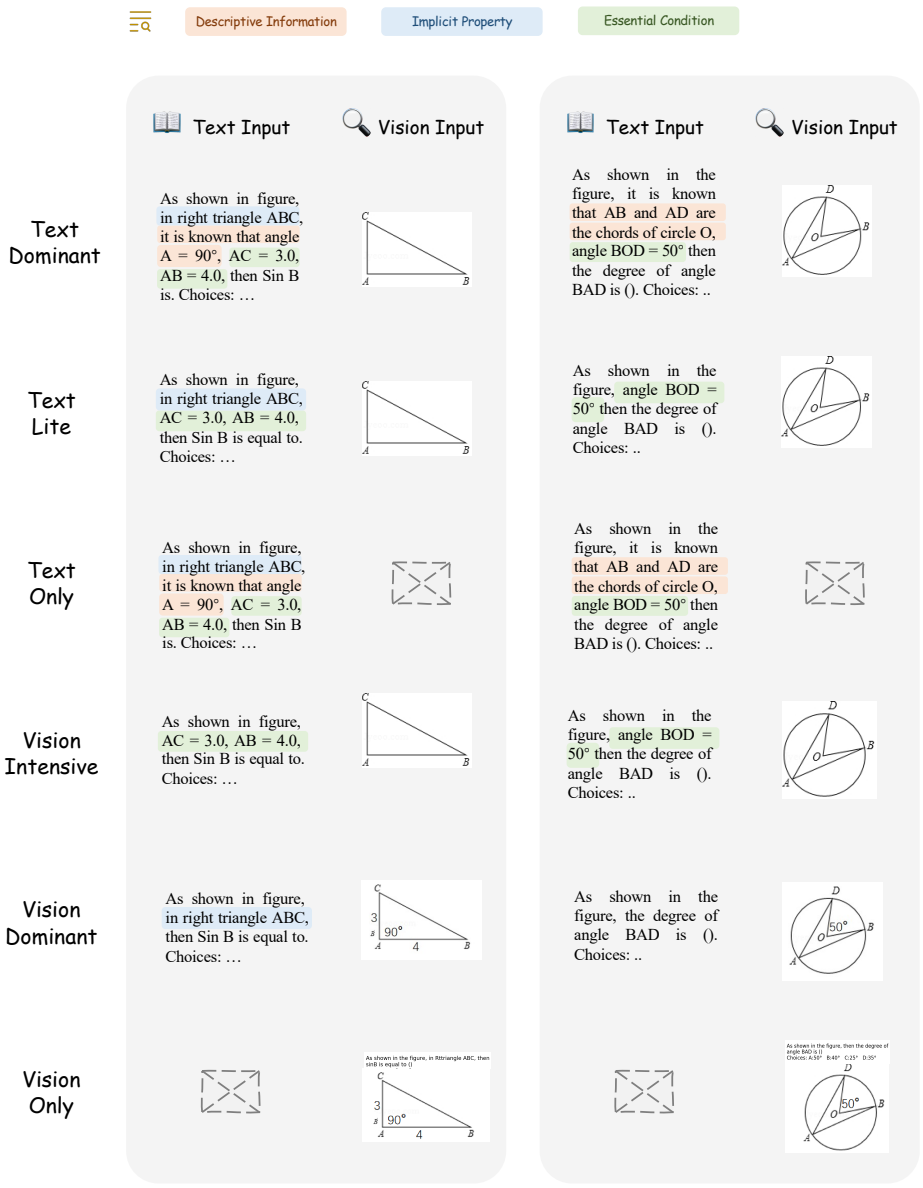
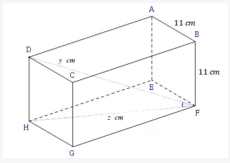


Fig. 15: Comparison of Six Problem Versions in MathVerse.



Question:

A square prism has a side GF of length 15 cm as shown. Now, we want to find y, the length of the diagonal DF. Calculate y to two decimal places.

Answer: y = 21.61

Subject: Solid Geometry

Subfield: Length

GPT-4V:

✓ 1/1 1. Consider the square prism as a combination of two right-angled triangles. Triangle DEF is one of them, and triangle DCF is the other.

✓ 1/1 2. Calculate the length of the diagonal EF of the bottom square face EFGH using the Pythagorean theorem. Since EF is the diagonal of a square of side length 15cm (side GF), we have: EF^2 = GF^2 + GE^2, EF^2 = 450. Now taking the square root to find EF: EF = √450 = 15√2 centimeters.

✗ 0/1 3. Find the diagonal DF which runs through the rectangular face DCEF. Triangle DEF is a right-angled triangle, hence we again apply Pythagorean theorem, with EF (just calculated) and DE being the legs, and DF being the hypotenuse. Since DE is equivalent in length to AB which is given as 11 cm, we have: DF^2 = DE^2 + EF^2 = 571.Taking the square root to find DF: DF = √571. Let's calculate the numeric value to two decimal places: DF ≈ 23.89 centimeters.

✗ 0/1 4. Therefore, y, the length of the diagonal DF, is approximately 23.89 cm.

Visual Perception Error

Multi-step Score: 0.5 Final CoT Evaluation Score: 0.35

LLaVA-NeXT:

✓ 1/1 1. Use the Pythagorean theorem to find the length of the diagonal. The theorem states that the square of the length of the hypotenuse (the diagonal in this case) is equal to the sum of the squares of the lengths of the legs. In this case, the legs are the sides of the square prism, which are of equal length.

✗ 0/1 2. Calculate the square of the lengths of the legs: y^2 = 11^2 + 11^2, y^2 = 121 + 121, y^2 = 242.

✗ 0/1 3. Take the square root of both sides to find the length of the diagonal: y = √242.

✗ 0/1 4. Round the square root to two decimal places and then multiply by itself to find the length of y to two decimal places: y = √242 ≈ 15.5, y * y ≈ 25.25.

✗ 0/1 5. The length of the diagonal DF is approximately 15.5 cm.

Numerical Calculation Error

Multi-step Score: 0.2 Final CoT Evaluation Score: 0.14

SPHINX-MoE:

✓ 1/1 1.Use the Pythagorean theorem to find the length of the diagonal DF.

✓ 1/1 2.The Pythagorean theorem states that in a right triangle, the square of the length of the hypotenuse (the side opposite the right angle) is equal to the sum of the squares of the lengths of the other two sides.

✗ 0/1 3.In this case, we have a right triangle with sides of length 11cm and 15cm. Let's call the length of the hypotenuse y.

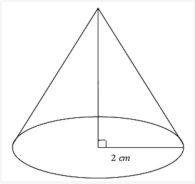
✗ 0/1 4. Using the Pythagorean theorem, we have: y^2 = 11^2 + 15^2, y^2 = 121 + 225, y^2 = 346.

✗ 0/1 5. Taking the square root of both sides, we get: y = √346, y ≈ 18.60.

✗ 0/1 6. Therefore, the length of the diagonal DF is approximately 18.60 cm.

Multi-step Score: 0.33 Final CoT Evaluation Score: 0.23

Fig. 16: Response Comparison of GPT-4V [46], LLaVA-NeXT [36], and SPHINX-MoE [19]. We adopt the Text-lite version of the problem, and highlight the key-step extraction and scoring by the CoT evaluation strategy.



Question:

Find the volume of the cone shown. The perpendicular height is 6 cm. Round your answer to two decimal places.

Answer: Volume = 25.13 cm^3

Subject: Solid Geometry

Subfield: Volume

GPT-4V:

✓ 1/1 1.Use the formula for the volume of a cone: Volume = 1/3 * π * r^2 * h, where r is the radius and h is the height of the cone.

✓ 1/1 2. Substitute the given values into the formula: Volume = 1/3 * π * (2 cm)^2 * 6 cm, Volume = 1/3 * π * 4 cm^2 * 6 cm, Volume = π * 4 cm^2 * 2 cm, Volume = 8 π cm^3.

✓ 1/1 3. Compute the volume numerically using $\pi \approx 3.14159$: Volume $\approx$ 8 * 3.14159, Volume $\approx$ 25.13272 cm^3.

✓ 1/1 4. Round the result to two decimal places: Volume $\approx$ 25.13 cm^3.

✓ 1/1 5. Final value: 25.13 cm^3.

Multi-step Score: 1 Final CoT Evaluation Score: 1

LLaVA-NeXT:

Reasoning Error

✓ 1/1 1. Use the formula for the volume of a cone: V = (1/3) π r^2 h, where r is the radius and h is the height of the cone.

✗ 0/1 2. Convert the height from centimeters to meters. There are 100 centimeters in a meter, so 6 centimeters is equal to 0.06 meters.

✗ 0/1 3. Plug the values into the formula: V = (1/3) π (2^2) (0.06), V = (1/3) π (4) (0.06), V = (1/3) π (4) (0.06) m^3, V = 0.00064 m^3.

✗ 0/1 4. Round to two decimal places. The volume of the cone is approximately 0.00064 cubic meters.

Multi-step Score: 0.25 Final CoT Evaluation Score: 0.175

SPHINX-MoE:

Knowledge Error

✓ 1/1 1. Use the formula for the volume of a cone: V = (1/3) π r^2 h, where V is the volume, r is the radius, and h is the height.

✗ 0/1 2. Given the height h = 6 cm but not the radius r. Use the fact that the radius is equal to the height of the triangle formed by the cone's slant height and the height of the cone.

✗ 0/1 3. The slant height is the hypotenuse of a right triangle with legs of length 6 cm and 2 cm (the radius of the base of the cone). Use the Pythagorean theorem to find the slant height: a^2 = b^2 + c^2, where a is the hypotenuse (slant height), b is the height of the cone (6 cm), and c is the radius of the base of the cone.

✗ 0/1 4. Calculate the slant height: a^2 = 6^2 + 2^2, a^2 = 36 + 4, a^2 = 40, a = $\sqrt{40}$.

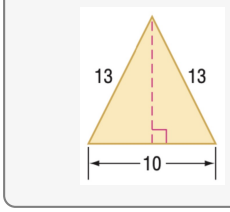
✗ 0/1 5. Use the Pythagorean theorem again to find the radius (r = 2 cm): r^2 = a^2 - b^2, r^2 = ($\sqrt{40}$)^2 - 6^2 = 40 - 36, r^2 = 4, r = $\sqrt{4}$.

✗ 0/1 6. Now that we have the radius (r = 2 cm), we can find the volume of the cone: V = (1/3) π r^2 h, V = (1/3) π (2)^2 (6), V = (1/3) π (4)(6), V = (1/3) π (24), V = 8π.

✓ 1/1 7. The volume of the cone is approximately 25.13 cubic centimeters.

Multi-step Score: 0.29 Final CoT Evaluation Score: 0.5

Fig. 17: Response Comparison of GPT-4V [46], LLaVA-NeXT [36], and SPHINX-MoE [19]. We adopt the Text-lite version of the problem, and highlight the key-step extraction and scoring by the CoT evaluation strategy.



Question:

There is an isosceles triangle. Find the area of the figure.
Choices: A:30 B:60 C:120 D:240

Answer: B

Subject: Plane Geometry

Subfield: Area

GPT-4V:

✓ 1/1 1. Use the formula for the area of a triangle, which is (1/2 * base * height).

✓ 1/1 2. Determine the height h of the isosceles triangle with two equal sides (each with length 13) and a base (length 10).

✓ 1/1 3. Drop a perpendicular from the vertex between the equal sides to the midpoint of the base to create two right-angled triangles. This splits the base into two segments of length 5 each.

✓ 1/1 4. Use the Pythagorean theorem for one of these right-angled triangles: h^2 + 5^2 = 13^2, h^2 = 13^2 - 5^2, h^2 = 169 - 25, h^2 = 144. So, h = √144. Therefore, h = 12.

✓ 1/1 5. Find the area of the isosceles triangle: Area = (1/2 * base * height), Area = (1/2 * 10 * 12), Area = 5 * 12, Area = 60.

✓ 1/1 6. The area of the isosceles triangle is 60. The correct option is B: 60.

Multi-step Score: 1 Final CoT Evaluation Score: 1

LLaVA-NeXT:

✓ 1/1 1. Calculate the area of an isosceles triangle using the formula: Area = (1/2) * Base * Height.

✓ 1/1 2. In the given problem, the base of the triangle is 10 units long, and the height is 13 units long.

✗ 0/1 3. Substitute the given values into the formula: Area = (1/2) * 10 * 13, Area = 13 * 7, Area = 91.

✗ 0/1 4. The final answer is 91.

Numerical Calculation Error

Multi-step Score: 0.5 Final CoT Evaluation Score: 0.35

SPHINX-MoE:

Visual Perception Error

✓ 1/1 1. The area of an isosceles triangle is given by the formula A = (1/2) * b * h, where b is the base length and h is the height.

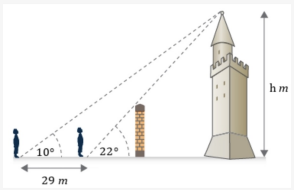
✗ 0/1 2. In this case, the base length is 10 and the height is 13.

✗ 0/1 3. Plugging these values into the formula, we get A = (1/2) * 10 * 13 = 65.

✗ 0/1 4. Therefore, the area of the triangle is 65. The correct option is B: 60.

Multi-step Score: 0.25 Final CoT Evaluation Score: 0.475

Fig. 18: Response Comparison of GPT-4V [46], LLaVA-NeXT [36], and SPHINX-MoE [19]. We adopt the Text-lite version of the problem, and highlight the key-step extraction and scoring by the CoT evaluation strategy.



Question:

Find the value of h to the nearest metre.

Answer: 9

Subject: Plane Geometry

Subfield: Applied

GPT-4V:

✓ 1/1 1. Use trigonometry, specifically the tangent function, to relate the opposite side of the angle in the triangle (which is h, the height of the tower) to the adjacent side (which is the horizontal distance from the base of the tower).

✓ 1/1 2. For the smaller angle of 10 degrees, the tower with the known height creates the opposite side, and the 29-meter distance is the adjacent side. However, we need to find the height of the taller tower, for which we have a 22-degree angle.

✓ 1/1 3. Use the tangent function for the 22-degree angle: tan(22°) = opposite/adjacent tan(22°) = h/29.

✓ 1/1 4. Solve for h: h = 29 * tan(22°).

✓ 1/1 5. Calculate this with a calculator: h ≈ 29 * 0.4040 (tan(22°) ≈ 0.4040) h ≈ 11.716 metres.

✓ 1/1 6. Round to the nearest metre: h ≈ 12 metres Final value: 12.

Multi-step Score: 1 Final CoT Evaluation Score: 1

LLaVA-NeXT:

✓ 1/1 1. Determine the height of the tower relative to the person's height by measuring the distance between the person's shadow and the tower.

✗ 0/1 2. Assume that the distance between the person and the tower is equal to the length of the shadow.

✓ 1/1 3. Use the formula to find the length of the shadow: Shadow length = (person's height) x tan(angle of light).

✗ 0/1 4. In this case, the angle of light is approximately 90 degrees, so the shadow length is: Shadow length = 2.2 meters x tan(90 degrees) = 2.2 meters x 1 = 2.2 meters.

Visual Perception Error

✗ 0/1 5. Find the height of the tower relative to the person's height by subtracting the shadow length from the tower's height: h = 2.5 meters - 2.2 meters = 0.3 meters.

✗ 0/1 6. The value of h is approximately 0.3 meters, or 30 centimeters.

Reasoning Error

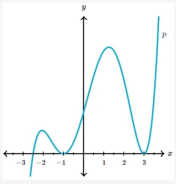
Multi-step Score: 0.3 Final CoT Evaluation Score: 0.2

SPHINX-MoE:

✗ 0/1 1. The single answer is that the value of h is 30.

Multi-step Score: 0 Final CoT Evaluation Score: 0

Fig. 19: Response Comparison of GPT-4V [46], LLaVA-NeXT [36], and SPHINX-MoE [19]. We adopt the Text-lite version of the problem, and highlight the key-step extraction and scoring by the CoT evaluation strategy.



Question:

A polynomial p is graphed. What could be the equation of p ?
Choices:
A: $p(x)=(x+1)^2(2x+5)^2(x-3)^2$
B: $p(x)=(x+1)^2(2x+5)(x-3)^2$
C: $p(x)=(x+1)^2(2x+5)(x-3)$
D: $p(x)=(x+1)(2x+5)(x-3)^2$

Answer: B **Subject:** Functions **Subfield:** Expression

GPT-4V:

✓ 1/1 1. Analyze the graph for key features such as x-intercepts (roots) and the behavior of the graph at those intercepts.

✓ 1/1 2. Identify the x-intercepts: at x = -1, x = -2.5 (which is -5/2 when converted to a fraction), and x = 3. This tells us that the factors of the polynomial must include (x + 1), (x + 5/2), and (x - 3).

✓ 1/1 3. Look at the behavior of the graph at each intercept to determine the multiplicity of each root. The graph touches the x-axis and turns around at x = -1 and x = 3, which indicates that these roots are of even multiplicity. The graph crosses the x-axis at x = -2.5, which indicates that this root is of odd multiplicity.

✓ 1/1 4. Match the graph to the correct equation by comparing the multiplicity of the roots with the given options.

✓ 1/1 5. The correct equation that matches the graph is: B: $p(x)=(x+1)^2(2x+5)(x-3)^2$.

Multi-step Score: 1 Final CoT Evaluation Score: 1

LLaVA-NeXT:

Visual Perception Error

✓ 1/1 1. Analyze the graph and identify the degree of the terms in the polynomial.

✗ 0/1 2. Identify that the polynomial has a degree of 2 for the terms with the highest degree, which are the squares of the linear terms.

✗ 0/1 3. Identify the highest degree term that is not squared is the term with the variable raised to the second power, which is $(x+1)^2$.

✗ 0/1 4. Identify the next highest degree term is $(2x+5)^2$, and the lowest degree term is $(x-3)^2$.

✗ 0/1 5. Conclude that the equation of the polynomial p(x) should be of the form: $p(x) = (x+1)^2(2x+5)^2(x-3)^2$ This is option A.

Multi-step Score: 0.2 Final CoT Evaluation Score: 0.14

SPHINX-MoE:

✗ 0/1 1. The graph of p has a vertical tangent at x=-1.

Visual Perception Error

✗ 0/1 2.The graph has a horizontal tangent at x=3.

✗ 0/1 3. The graph has a vertical tangent at x=-5/2.

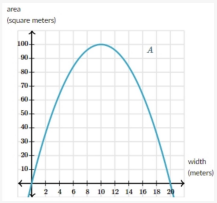
✓ 1/1 4.The graph of p also passes through the point (-1,0).

✗ 0/1 5. The equation of p is $p(x)=(x+1)^2(2\ x+5)(x-3)^2$.

✗ 0/1 6. The correct option is A.

Multi-step Score: 0.17 Final CoT Evaluation Score: 0.12

Fig. 20: Response Comparison of GPT-4V [46], LLaVA-NeXT [36], and SPHINX-MoE [19]. We adopt the Text-lite version of the problem, and highlight the key-step extraction and scoring by the CoT evaluation strategy.



Question:

The function A models the rectangle's area (in square meters) as a function of its width (in meters). Which of these statements are true? Choices:

A: Greater width always relates to smaller area.
B: Greater width relates to smaller area as long as the width is less than 10 m.
C: When there is no width, the area is 20 m^2.
D: When there is no width, the area is 0 m^2.

Answer: D **Subject:** Functions **Subfield:** Applied

GPT-4V:

✓ 1/1 1.Analyze the graph of the function A, which models the area of a rectangle as a function of its width.

✓ 1/1 2. Evaluate each statement based on the graph:
- A: Greater width always relates to a smaller area.
- B: The area is always increasing with the width.
- C: When the width is 0, the area is not 0.
- D: When there is no width, the area is 0 m^2.

✓ 1/1 3. Determine that statement D is true because the graph starts at the origin (0,0), indicating that when there is no width, there is no area, which is consistent with the properties of a rectangle.

✓ 1/1 4. The correct option is: D: When there is no width, the area is 0 m^2.

Multi-step Score: 1 Final CoT Evaluation Score: 1

LLaVA-NeXT:

Visual Perception Error

✓ 1/1 1. Analyze the given information and the graph of the quadratic function A.

✗ 0/1 2. Observe that the vertex of the quadratic function is at (10, 100), indicating that the function has a minimum point at this point.

✓ 1/1 3. Note that the quadratic function intersects the x-axis at (0,0) and (20,0), which means the area is zero when the width is zero.

✓ 1/1 4. Consider the given statements and determine their correctness based on the observations from the graph.

✓ 1/1 5. Statement A is incorrect as the area does not always decrease with increasing width.

✗ 0/1 6. Statement B is correct as the area decreases with increasing width up to the vertex, which is at 10 meters.

✓ 1/1 7. Statement C is incorrect as the area is zero when the width is zero, not 20 square meters.

✓ 1/1 8. Statement D is correct as the area is zero when the width is zero.

✗ 0/1 9. The correct statement is B: Greater width relates to smaller area as long as the width is less than 10 m.

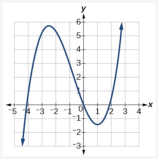
Multi-step Score: 0.67 Final CoT Evaluation Score: 0.47

SPHINX-MoE:

✓ 1/1 1. The single answer is D.

Multi-step Score: 1 Final CoT Evaluation Score: 1

Fig. 21: Response Comparison of GPT-4V [46], LLaVA-NeXT [36], and SPHINX-MoE [19]. We adopt the Text-lite version of the problem, and highlight the key-step extraction and scoring by the CoT evaluation strategy.



Question: (Text Dominant)

There is a curve goes through (-4,1), (-1,3), and (0,0). Estimate the intervals on which the polynomial function is increasing or decreasing.

Answer: The function is increasing on $(-\infty,-2.5) \cup (1, \infty)$, and decreasing on (-2.5,1)

Subject: Functions

Subfield: Property

Text Dominant:

✓ 1/1 1. Determine the intervals on which the function is increasing or decreasing by looking at the slope of the curve on the graph.

✗ 0/1 2. The function starts off decreasing from the point (-5, something above 6) to the point (-4, 1).

✗ 0/1 3. The function then starts increasing at the point (-4, 1) and continues to increase until it reaches a peak somewhere between -2 and -1 on the x-axis.

✗ 0/1 4. After reaching the peak, the function starts to decrease again, and it continues to decrease past the point (-1, 3) until it reaches the point (0, 0).

✗ 0/1 5. The function reaches its minimum at (0, 0) and then starts to increase again. It continues to increase past the point (2, something below -3) and keeps increasing as it moves off to the right of the graph.

✗ 0/1 6. The intervals on which the function is increasing or decreasing are as follows:
- Decreasing: from the start of the graph to x = -4, and from just before x = -2 to x = 0.
- Increasing: from x = -4 to just before x = -2, and from x = 0 to the end of the graph.

✗ 0/1 7. More precisely:
- The function is increasing on the intervals (-4, -1.5) and (0, ∞).
- The function is decreasing on the intervals (-∞, -4) and (-1.5, 0).

✓ 1/1 8. Note that the exact turning points are estimated since the graph does not provide precise values, and the intervals are given in terms of x-values. The exact values of the turning points cannot be determined without more information or a more precise graph.

Multi-step Score: 0.25 Final CoT Evaluation Score: 0.175

Text Lite:

✓ 1/1 1. Determine the slope of the function as it moves from left to right along the x-axis.

✓ 1/1 2. Identify the points where the function reaches its peaks and its lowest points.

✗ 0/1 3. The function is increasing from the leftmost part of the graph until it reaches its first peak, which is approximately at x = -3.

✗ 0/1 4. The function starts decreasing after this peak until it reaches its lowest point, which is approximately at x = 0.

✗ 0/1 5. The function then starts increasing again from x = 0 until it reaches its next peak, which is approximately at x = 2.

✗ 0/1 6. The function decreases again after x = 2 towards the rightmost part of the graph.

✗ 0/1 7. The intervals on which the function is increasing are from x = -5 to x = -3 and from x = 0 to x = 2. The intervals on which the function is decreasing are from x = -3 to x = 0 and from x = 2 to x = 4.

✗ 0/1 8. Final values:
Increasing intervals: (-5, -3) and (0, 2),
Decreasing intervals: (-3, 0) and (2, 4).

Multi-step Score: 0.25 Final CoT Evaluation Score: 0.175

Vision Dominant:

✓ 1/1 1. Determine the intervals on which the function is increasing or decreasing by looking at the slope of the function as it moves from left to right along the x-axis.

✗ 0/1 2. From the graph, identify that the function is increasing from the leftmost part of the graph until it reaches its first peak. This occurs from approximately x = -5 to x = -2.5.

✗ 0/1 3. After reaching the first peak, the function starts to decrease until it reaches the bottom of the first trough. This occurs from approximately x = -2.5 to x = 0.

✗ 0/1 4. The function then starts to increase again from the bottom of the trough until it reaches the second peak. This occurs from approximately x = 0 to x = 1.

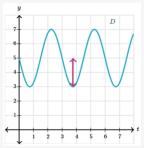
✗ 0/1 5. Finally, the function decreases again after the second peak towards the rightmost part of the graph. This occurs from approximately x = 1 to x = 4.

✗ 0/1 6. Conclude that the function is increasing on the intervals: (-5, -2.5) and (0, 1)

✗ 0/1 7. And the function is decreasing on the intervals: (-2.5, 0) and (1, 4)

Multi-step Score: 0.14 Final CoT Evaluation Score: 0.1

Fig. 22: Response of Different Problem Versions by GPT-4V [46]. By comparing the output from three problem versions, we observe increasing reasoning errors when more textual content is visually depicted in diagrams. Note that we only showcase the Text-dominant version of the problem for briefness.



Question: (Text Dominant)

Sia is swinging from a chandelier. The horizontal distance between Sia and the wall, in meters, is modeled by D(t) where t is the time in seconds. The function is graphed below, along with one segment highlighted, a red double arrow from (3.75,3), (3.75,5). What is the meaning of the highlighted segment?

Choices:

A:The chandelier is hung 2 meters from the wall.

B:The furthest Sia gets from the point where the chandelier hangs is 2 meters.

C:The chandelier is 2 meters wide.

D:Sia completes 2 swing cycles per second.

Answer: B **Subject:** Functions **Subfield:** Applied

Text Dominant:

✓ 1/1 1. The graph shows a function D(t) that represents the horizontal distance between Sia and the wall over time.

✓ 1/1 2. The highlighted segment with a red double arrow spans vertically from the point (3.75, 3) to the point (3.75, 5) on the graph.

✗ 0/1 3. This segment is at a constant time of 3.75 seconds and shows a change in the distance D(t) from 3 meters to 5 meters.

✗ 0/1 4. The segment does not provide information about the width of the chandelier or the number of swing cycles per second. Instead, it indicates a change in distance at a specific moment in time.

✓ 1/1 5. Since the distance changes from 3 to 5 meters, this represents a 2-meter difference in the horizontal distance from the wall during the swing.

✓ 1/1 6. Therefore, the meaning of the highlighted segment is that the furthest Sia gets from the point where the chandelier hangs is 2 meters.

✓ 1/1 7. The correct option is: B: The furthest Sia gets from the point where the chandelier hangs is 2 meters.

Multi-step Score: 0.71 Final CoT Evaluation Score: 0.8

Text Lite:

✓ 1/1 1. Analyze the graph which shows a periodic function, representing the back-and-forth motion of Sia swinging from the chandelier. The horizontal axis is labeled with time t in seconds, and the vertical axis is labeled with distance D in meters. The highlighted segment on the graph indicates a change in the distance from one point to another over time.

✓ 1/1 2. Evaluate each choice:

✓ 1/1 - A: The chandelier is hung 2 meters from the wall. This statement would imply a fixed distance from the wall, which the graph does not provide information about.

✗ 0/1 - B: The furthest Sia gets from the point where the chandelier hangs is 2 meters. This statement seems to be about the amplitude of the swing, but the highlighted segment does not indicate the amplitude; it indicates a change in distance over time.

✓ 1/1 - C: The chandelier is 2 meters wide. The width of the chandelier is not something that can be determined from this graph, which shows distance from the wall over time.

✓ 1/1 - D: Sia completes 2 swing cycles per second. The highlighted segment shows a change in distance over time, but it does not provide information about the number of cycles per second.

✗ 0/1 3. Conclude that none of the options A, B, C, or D directly describe the meaning of the highlighted segment in terms of the graph's representation of distance over time.

✗ 0/1 4. Given the options provided and the information from the graph, none of the choices accurately describe the meaning of the highlighted segment. Therefore, based on the information provided and the reasoning above, there is no correct option to choose from A, B, C, or D.

Multi-step Score: 0.63 Final CoT Evaluation Score: 0.44

Vision Dominant:

✓ 1/1 1. The graph shows a periodic function, representing the motion of a pendulum or a swing. The horizontal axis represents time (t), and the vertical axis represents the horizontal distance (D) from the wall. The highlighted segment indicates a change in the distance D over time.

✗ 0/1 2. The highlighted segment is a vertical line, showing a change in distance at a particular moment in time, from a lower value to a higher value on the D(t) function.

✗ 0/1 3. None of the options A, B, C, or D directly relate to the meaning of a single vertical segment on the graph. The closest interpretation of the highlighted segment would be the change in distance from the wall at a specific moment, which is not explicitly described in any of the options provided.

✗ 0/1 4. Given the information and the graph, none of the options A, B, C, or D correctly describe the meaning of the highlighted segment.

Multi-step Score: 0.25 Final CoT Evaluation Score: 0.175

Fig. 23: Response of Different Problem Versions by GPT-4V [46]. By comparing the output from three problem versions, we observe increasing reasoning errors when more textual content is visually depicted in diagrams. Note that we only showcase the Text-dominant version of the problem for briefness.

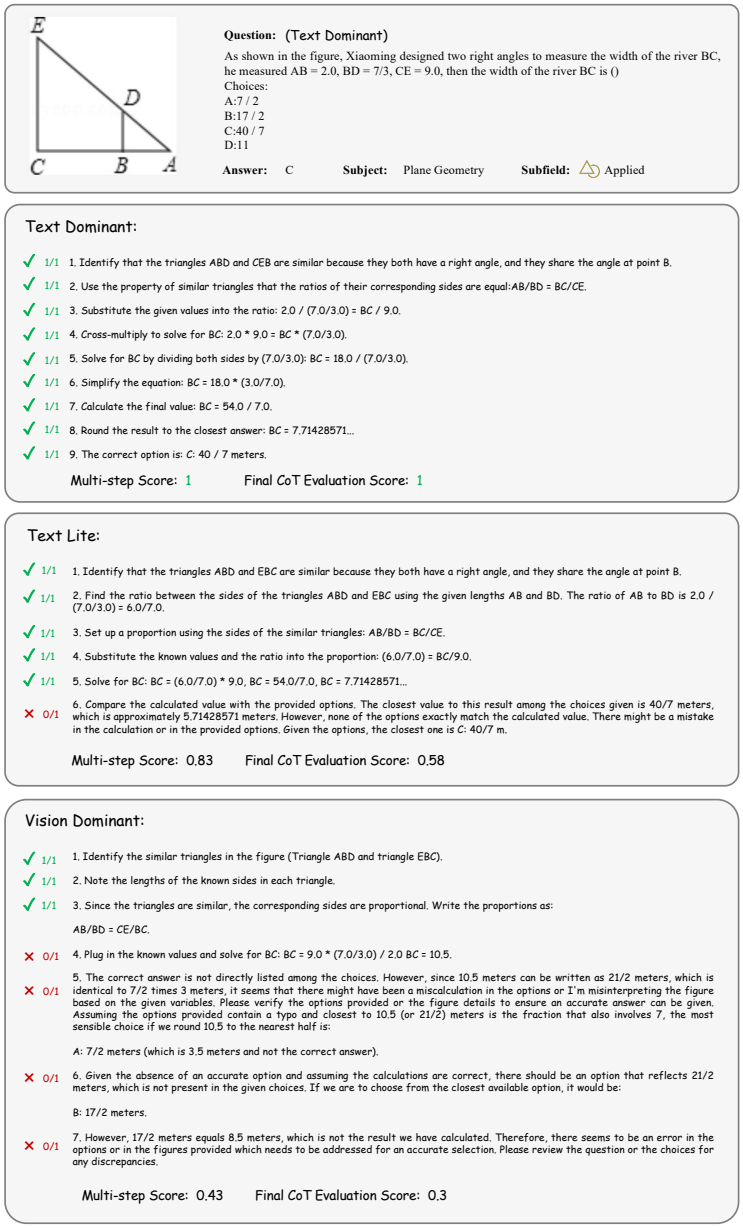
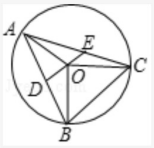


Fig. 24: Response of Different Problem Versions by GPT-4V [46]. By comparing the output from three problem versions, we observe increasing reasoning errors when more textual content is visually depicted in diagrams. Note that we only showcase the Text-dominant version of the problem for briefness.



Question: (Text Dominant)

As shown in the figure, triangle ABC is inscribed in circle O, OC perpendicular OB, OD perpendicular AB intersects AC at point E. Knowing that the radius of circle O is 1.0, then what is the value of $AE^2 + CE^2$?

Choices:

A:1
B:2
C:3
D:4

Answer: B **Subject:** Plane Geometry **Subfield:** Length

Text Dominant:

✓ 1/1 1. Identify that OC is perpendicular to OB, making triangle OCB a right triangle with OC and OB as its legs. Since OC and OB are radii of the circle and the radius is given as 1.0, OC = OB = 1.0. Therefore, triangle OCB is an isosceles right triangle.

✓ 1/1 2. Identify that OD is perpendicular to AB, which means that OD is the altitude of triangle AOB from point O to side AB. Since triangle AOB is also isosceles (OA = OB = 1.0), OD bisects AB. This means that AD = DB.

✓ 1/1 3. Consider triangle AOD and triangle BOD. Since OD is perpendicular to AB, both triangles AOD and BOD are right triangles. Also, since AD = DB, triangle AOD is congruent to triangle BOD. Therefore, AE = EB.

✗ 0/1 4. Since AE = EB and AC = AB (because triangle ABC is inscribed in the circle and AC and AB are both radii), we can say that CE = EB.

✗ 0/1 5. To find AE^2 + CE^2, since AE = CE, we can say that AE^2 + CE^2 = 2 * AE^2.

✗ 0/1 6. To find AE, use the Pythagorean theorem in triangle AOD (or BOD, since they are congruent). We have:

AO^2 = AD^2 + OD^2.

Since AO = 1.0 (radius of the circle) and AD = 1/2 * AB (because AD = DB and AB is the diameter of the circle, which is twice the radius), we have:

1.0^2 = (1/2 * AB)^2 + OD^2.

AB is the diameter of the circle, so AB = 2 * radius = 2 * 1.0 = 2.0. Therefore: 1.0^2 = (1/2 * 2.0)^2 + OD^2 OD^2 = 0
Since OD^2 = 0, OD must be 0, which means that point D coincides with point O, and AE = EB = 1.0 (since AE = EB and AB = 2 * radius = 2.0). Therefore, AE^2 + CE^2 = 2 * AE^2 = 2 * 1.0^2 = 2.

✓ 1/1 7. The correct answer is: B: 2

Multi-step Score: 0.57 Final CoT Evaluation Score: 0.7

Text Lite:

✓ 1/1 1. Use the properties of a circle and the Pythagorean theorem.

✓ 1/1 2. Since OC is perpendicular to OB and both are radii of the circle, triangle OCB is an isosceles right triangle. Therefore, the length of BC can be found using the Pythagorean theorem: BC = √2

✓ 1/1 3. Similarly, since OD is perpendicular to AB, triangle OAD is also an isosceles right triangle. Therefore, AD = OD = 1.

✗ 0/1 4. Using the Pythagorean theorem again, AO = √2. Since AO is the diameter of the circle, AC is also the diameter, and thus AC = AO = √2.

✗ 0/1 5. We have AC = √2 and BC = √2. Since E is a point on AC, AE + EC = AC.

✗ 0/1 6. We want to find AE^2 + CE^2. Let's denote AE as x and CE as (√2 - x). Then we can express AE^2 + CE^2 as: 2x^2 - 2√2x + 2.

✗ 0/1 7. To minimize this expression, we can take the derivative with respect to x and set it to zero to find the minimum value. However, since we are given that AE and CE are segments of a line, and we know that the sum of their squares in a right triangle is constant, we can infer that the sum of the squares of AE and CE will be equal to the square of AC.

✓ 1/1 8. Therefore, AE^2 + CE^2 = AC^2 = (√2)^2 = 2.

✓ 1/1 9. The correct answer is: B: 2

Multi-step Score: 0.56 Final CoT Evaluation Score: 0.69

Vision Dominant:

✓ 1/1 1. Given that the radius of circle O is 1, this means that OA, OE, and OC are all radii of the circle, and thus they all have a length of 1.

✓ 1/1 2. Use the Pythagorean theorem (a^2 + b^2 = c^2) on triangle AOE, where AE is the hypotenuse, and OA and OE are the two legs:

AE^2 = OA^2 + OE^2, AE^2 = 2.

✓ 1/1 3. By the same logic, apply the Pythagorean theorem to triangle COE, where CE is the hypotenuse:

CE^2 = OC^2 + OE^2, CE^2 = 2.

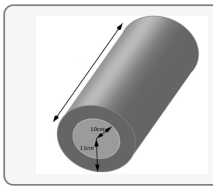
✗ 0/1 4. Now, according to the question, find the value of AE^2 + CE^2:

AE^2 + CE^2 = 2 + 2 AE^2 + CE^2 = 4.

✗ 0/1 5. Therefore, the value of AE^2 + CE^2 is 4. The correct option is D: 4.

Multi-step Score: 0.6 Final CoT Evaluation Score: 0.42

Fig. 25: Response of Different Problem Versions by GPT-4V [46]. By comparing the output from three problem versions, we observe increasing reasoning errors when more textual content is visually depicted in diagrams. Note that we only showcase the Text-dominant version of the problem for briefness.



Question: (Text Dominant)

Consider the solid pictured and answer the following: Hence what is the total surface area? The outer radius is 11 cm, inner radius is 10 cm and length 24 cm. Give your answer to the nearest two decimal places.

Answer: SA = 3298.67 cm^2

Subject: Solid Geometry

Subfield: Area

Text Dominant

✓ 1/1 1. Calculate the surface area of the outer side of the cylinder using the formula for the lateral surface area of a cylinder: $2\pi rh$, where r is the outer radius and h is the length (height) of the cylinder. Outer surface area = $2\pi * 11 * 24$.

✓ 1/1 2. Calculate the surface area of the inner side of the cylinder using the same formula, but with the inner radius. Inner surface area = $2\pi * 10 * 24$.

✓ 1/1 3. Calculate the surface area of both circular ring-shaped ends. This is the area of the outer circle minus the area of the inner circle, multiplied by 2 (since there are two ends). Area for one end = Area of outer circle - Area of inner circle = πr_{outer}^2 - πr_{inner}^2 = π11^2 - π10^2. Total area for both ends = 2 * (π(11^2 - 10^2)).

✗ 0/1 4. Add all these together to get the total surface area: Total surface area = $528\pi + 480\pi + 42\pi = 1050\pi$.

✓ 1/1 5. Use the approximate value of π, which is 3.14, to calculate the total surface area in decimal form: Total surface area ≈ 1050 * 3.14.

✗ 0/1 6. Round the result to the nearest two decimal places: 3297.00.

✗ 0/1 7. The total surface area of the solid, to the nearest two decimal places, is approximately 3297.00 square centimeters.

Multi-step Score: 0.57 Final CoT Evaluation Score: 0.4

Text Lite

✗ 0/1 1. Calculate the area A of a circle using the formula A = πr^2, where r is the radius. In this case, the radius r is 5 cm. Therefore, the area of one circular base is π(5^2) or 25π.

✗ 0/1 2. Since a cylinder has two bases (top and bottom), multiply the area of one base by 2 to find the total area of both bases: $2 \times 25\pi = 50\pi$ cm^2.

✗ 0/1 3. To find the area of the curved surface of the cylinder, use the formula for the lateral surface area of a cylinder, which is A = $2\pi rh$, where h is the height or length of the cylinder. In this case, the height h is 24 cm. The lateral surface area would therefore be $2\pi rh = 2\pi(5)(24) = 240\pi$ cm^2.

✗ 0/1 4. Add both calculated areas together to find the total surface area: Total surface area = area of both bases + lateral surface area. Total surface area = $50\pi + 240\pi = 290\pi$ cm^2.

✗ 0/1 5. Use the approximation $\pi \approx 3.14159$ to compute the numerical answer: Total surface area ≈ 290×3.14159. Total surface area ≈ 910.8611 cm^2.

✗ 0/1 6. Rounded to the nearest two decimal places, the total surface area is 910.86 cm^2 .

Multi-step Score: 0 Final CoT Evaluation Score: 0

Vision Dominant

✗ 0/1 1. Find the area of one circular base using the formula A = πr^2, where r is the radius of the base. In this case, the radius is 5 cm, so the area of one base is $\pi * 25$ cm^2 = 78.54 cm^2 (to two decimal places).

✗ 0/1 2. Since there are two circular bases, double the result to get the total area of the bases: 2 * 78.54 cm^2 = 157.08 cm^2.

✗ 0/1 3. Calculate the rectangular surface area, which is the curved surface that wraps around the cylinder. The length of the rectangle is the same as the circumference of the base circle, which can be found with the formula $C = 2\pi r$. In this case, the circumference is $2 * \pi * 5$ cm = 31.42 cm (to two decimal places).

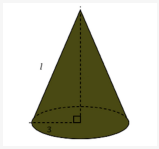
✗ 0/1 4. The width of the rectangle is the same as the height (length) of the cylinder, which is given as 24 cm. So the area of the rectangle is 31.42 cm * 24 cm = 753.96 cm^2 (to two decimal places).

✗ 0/1 5. Add the areas of both circles and the rectangular wrap to get the total surface area: 157.08 cm^2 + 753.96 cm^2 = 911.04 cm^2.

✗ 0/1 6. The total surface area of the cylinder to the nearest two decimal places is 911.04 cm^2 .

Multi-step Score: 0 Final CoT Evaluation Score: 0

Fig. 26: Response of Different Problem Versions by GPT-4V [46]. By comparing the output from three problem versions, we observe increasing reasoning errors when more textual content is visually depicted in diagrams. Note that we only showcase the Text-dominant version of the problem for briefness.



Question: (Text Dominant)

Find the volume of the cone pictured. The perpendicular height is 6 and radius is 3. (Give your answer correct to 2 decimal places.)

Answer: 56.55

Subject: Solid Geometry

Subfield: Volume

Text Dominant

✓ 1/1 1. Use the formula for the volume of a cone: V = 1/3πr^2h , where V is the volume, r is the radius of the base, and h is the height of the cone.

✓ 1/1 2. Given the radius r = 3 and the height h = 6, plug these values into the formula: V = 1/3π(3)^2 (6) , V = 1/3 π * 9 * 6 , V = 18 π.

✓ 1/1 3. Calculate the numerical value using the approximation π ≈ 3.14159: V ≈ 18 * 3.14159 , V ≈ 56.54862 π.

✓ 1/1 4. Round to two decimal places, the volume of the cone is approximately: V ≈ 56.55.

✓ 1/1 5. Final Value: 56.55.

Multi-step Score: 1 Final CoT Evaluation Score: 1

Text Lite

✓ 1/1 1. Use the formula for the volume of a cone: Volume = 1/3 π r^2 h, where r is the radius of the base of the cone, and h is the height.

✓ 1/1 2. Assume the "3" refers to the radius of the base of the cone, and the perpendicular height provided in the question text is "6". These values are the radius and height respectively.

✓ 1/1 3. Substitute r = 3 and h = 6 into the formula: Volume = 1/3 π (3)^2 (6), Volume = 1/3 π (9) (6), Volume = 1/3 π (54), Volume = 18 π.

✓ 1/1 4. Calculate the numeric value using the approximation π ≈ 3.14159: Volume = 18 * 3.14159, Volume ≈ 56.54862.

✓ 1/1 5. Round to two decimal places, the volume of the cone is: Volume ≈ 56.55.

✓ 1/1 6. Final value: 56.55.

Multi-step Score: 1 Final CoT Evaluation Score: 1

Vision Dominant

✓ 1/1 1. Identify the shape of the solid figure. In this case, it is a cone.

✓ 1/1 2. Use the formula for the volume of a cone: V = 1/3πr^2 h, where V is the volume, r is the radius of the base, h is the height of the cone, and \pi is the constant approximately equal to 3.14159.

✗ 0/1 3. Identify the values for r and h from the image. Here, h = 6 and the diameter of the base is 3, which means the radius r = 3/2 = 1.5.

✗ 0/1 4. Substitute the values into the formula: V = 1/3π(1.5)^2 * 6, V = 1/3π* 2.25 * 6, V = 1/3π* 13.5, V = π * 4.5, V ≈ 3.14159 * 4.5, V ≈ 14.13765.

✓ 1/1 5. Round the final answer to two decimal places. The volume is 14.14 cubic units.

✗ 0/1 6. Final answer: 14.14.

Multi-step Score: 0.5 Final CoT Evaluation Score: 0.35

Fig. 27: Response of Different Problem Versions by GPT-4V [46]. By comparing the output from three problem versions, we observe increasing reasoning errors when more textual content is visually depicted in diagrams. Note that we only showcase the Text-dominant version of the problem for briefness.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
2. Amini, A., Gabriel, S., Lin, P., Koncel-Kedziorski, R., Choi, Y., Hajishirzi, H.: Mathqa: Towards interpretable math word problem solving with operation-based formalisms. arXiv preprint arXiv:1905.13319 (2019)
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: Advances in neural information processing systems. pp. 1877–1901 (2020)
5. Cao, J., Xiao, J.: An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In: Proceedings of the 29th International Conference on Computational Linguistics. pp. 1511–1520 (2022)
6. Chen, G., Zheng, Y.D., Wang, J., Xu, J., Huang, Y., Pan, J., Wang, Y., Wang, Y., Qiao, Y., Lu, T., et al.: Videollm: Modeling video sequence with large language models. arXiv preprint arXiv:2305.13292 (2023)
7. Chen, J., Li, T., Qin, J., Lu, P., Lin, L., Chen, C., Liang, X.: Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. ArXiv **abs/2212.02746** (2022)
8. Chen, J., Li, T., Qin, J., Lu, P., Lin, L., Chen, C., Liang, X.: Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. arXiv preprint arXiv:2212.02746 (2022)
9. Chen, J., Tang, J., Qin, J., Liang, X., Liu, L., Xing, E.P., Lin, L.: Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. ArXiv **abs/2105.14517** (2021), `https://api.semanticscholar.org/CorpusID:235253782`
10. Chen, J., Tang, J., Qin, J., Liang, X., Liu, L., Xing, E.P., Lin, L.: Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. arXiv preprint arXiv:2105.14517 (2021)
11. Chen, J., Li, D.Z.X.S.X., Zhang, Z.L.P., Xiong, R.K.V.C.Y., Elhoseiny, M.: Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478 (2023)
12. Chen, L., Li, J., wen Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. ArXiv **abs/2311.12793** (2023), `https://api.semanticscholar.org/CorpusID:265308687`
13. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al.: Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168 (2021)
14. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023)
15. Dong, X., Zhang, P., Zang, Y., Cao, Y., Wang, B., Ouyang, L., Wei, X., Zhang, S., Duan, H., Cao, M., et al.: Internlm-xcomposer2: Mastering free-form text-image

composition and comprehension in vision-language large model. arXiv preprint arXiv:2401.16420 (2024)
16. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)
17. Gao, J., Pi, R., Zhang, J., Ye, J., Zhong, W., Wang, Y., Hong, L., Han, J., Xu, H., Li, Z., et al.: G-llava: Solving geometric problem with multi-modal large language model. arXiv preprint arXiv:2312.11370 (2023)
18. Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., Li, H., Qiao, Y.: Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010 (2023)
19. Gao, P., Zhang, R., Liu, C., Qiu, L., Huang, S., Lin, W., Zhao, S., Geng, S., Lin, Z., Jin, P., et al.: Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. arXiv preprint arXiv:2402.05935 (2024)
20. Gemini Team, G.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
21. Guo, Z., Zhang, R., Chen, H., Gao, J., Gao, P., Li, H., Heng, P.A.: Sciverse. https://sciverse-cuhk.github.io (2024), `https://sciverse-cuhk.github.io/`
22. Guo, Z., Zhang, R., Zhu, X., Tang, Y., Ma, X., Han, J., Chen, K., Gao, P., Li, X., Li, H., et al.: Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. arXiv preprint arXiv:2309.00615 (2023)
23. Han, J., Zhang, R., Shao, W., Gao, P., Xu, P., Xiao, H., Zhang, K., Liu, C., Wen, S., Guo, Z., et al.: Imagebind-llm: Multi-modality instruction tuning. arXiv preprint arXiv:2309.03905 (2023)
24. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. Proceedings of the International Conference on Learning Representations (ICLR) (2021)
25. Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., Steinhardt, J.: Measuring mathematical problem solving with the math dataset. NeurIPS (2021)
26. Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., Gan, C.: 3d-llm: Injecting the 3d world into large language models. Advances in Neural Information Processing Systems **36** (2024)
27. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., de Las Casas, D., Hanna, E.B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L.R., Saulnier, L., Lachaux, M., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T.L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mixtral of experts. Arxiv 2401.04088 (2024)
28. Jiang, D., Song, G., Wu, X., Zhang, R., Shen, D., Zong, Z., Liu, Y., Li, H.: Comat: Aligning text-to-image diffusion model with image-to-text concept matching. arXiv preprint arXiv:2404.03653 (2024)
29. Li, B., Zhang, Y., Chen, L., Wang, J., Pu, F., Yang, J., Li, C., Liu, Z.: Mimic-it: Multi-modal in-context instruction tuning. arXiv preprint arXiv:2306.05425 (2023)
30. Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension. ArXiv **abs/2307.16125** (2023)
31. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895 (2024)

32. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
33. Li, X., Zhang, M., Geng, Y., Geng, H., Long, Y., Shen, Y., Zhang, R., Liu, J., Dong, H.: Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. CVPR 2024 (2023)
34. Lin, Z., Liu, C., Zhang, R., Gao, P., Qiu, L., Xiao, H., Qiu, H., Lin, C., Shao, W., Chen, K., et al.: Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. arXiv preprint arXiv:2311.07575 (2023)
35. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)
36. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/
37. Liu, J., Li, C., Wang, G., Lee, L., Zhou, K., Chen, S., Xiong, C., Ge, J., Zhang, R., Zhang, S.: Self-corrected multimodal large language model for end-to-end robot manipulation. arXiv preprint arXiv:2405.17418 (2024)
38. Liu, J., Liu, M., Wang, Z., Lee, L., Zhou, K., An, P., Yang, S., Zhang, R., Guo, Y., Zhang, S.: Robomamba: Multimodal state space model for efficient robot reasoning and manipulation. arXiv preprint arXiv:2406.04339 (2024)
39. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? arXiv preprint arXiv:2307.06281 (2023)
40. Lu, P., Bansal, H., Xia, T., Liu, J., yue Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. ArXiv **abs/2310.02255** (2023)
41. Lu, P., Gong, R., Jiang, S., Qiu, L., Huang, S., Liang, X., Zhu, S.C.: Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In: Annual Meeting of the Association for Computational Linguistics (2021), https://api.semanticscholar.org/CorpusID:234337054
42. Lu, P., Gong, R., Jiang, S., Qiu, L., Huang, S., Liang, X., Zhu, S.C.: Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. arXiv preprint arXiv:2105.04165 (2021)
43. Luo, H., Sun, Q., Xu, C., Zhao, P., Lou, J., Tao, C., Geng, X., Lin, Q., Chen, S., Zhang, D.: Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. arXiv preprint arXiv:2308.09583 (2023)
44. OpenAI: Chatgpt. https://chat.openai.com (2023)
45. OpenAI: Gpt-4 technical report. ArXiv **abs/2303.08774** (2023)
46. OpenAI: GPT-4V(ision) system card (2023), https://openai.com/research/gpt-4v-system-card
47. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Gray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. In: Advances in Neural Information Processing Systems (2022)
48. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021), https://api.semanticscholar.org/CorpusID:231591445

49. Roy, S., Roth, D.: Solving general arithmetic word problems. ArXiv **abs/1608.01413** (2016), https://api.semanticscholar.org/CorpusID:560565
50. Seo, M., Hajishirzi, H., Farhadi, A., Etzioni, O., Malcolm, C.: Solving geometry problems: Combining text and diagram interpretation. In: Proceedings of the 2015 conference on empirical methods in natural language processing. pp. 1466–1476 (2015)
51. Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., Cai, D.: Pandagpt: One model to instruction-follow them all. arXiv preprint arXiv:2305.16355 (2023)
52. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
53. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
54. Wang, K., Ren, H., Zhou, A., Lu, Z., Luo, S., Shi, W., Zhang, R., Song, L., Zhan, M., Li, H.: Mathcoder: Seamless code integration in LLMs for enhanced mathematical reasoning. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=z8TW0ttBPp
55. Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., Meng, F., Huang, S., Qiao, Y., Luo, P.: Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. arXiv preprint arXiv:2306.09265 (2023)
56. Xu, R., Wang, X., Wang, T., Chen, Y., Pang, J., Lin, D.: Pointllm: Empowering large language models to understand point clouds. arXiv preprint arXiv:2308.16911 (2023)
57. Yang, S., Liu, J., Zhang, R., Pan, M., Guo, Z., Li, X., Chen, Z., Gao, P., Guo, Y., Zhang, S.: Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. arXiv preprint arXiv:2312.14074 (2023)
58. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., Jiang, C., Li, C., Xu, Y., Chen, H., Tian, J., Qian, Q., Zhang, J., Huang, F.: mplug-owl: Modularization empowers large language models with multimodality (2023)
59. Ye, Q., Xu, H., Ye, J., Yan, M., Hu, A., Liu, H., Qian, Q., Zhang, J., Huang, F., Zhou, J.: mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration (2023)
60. Ying, K., Meng, F., Wang, J., Li, Z., Lin, H., Yang, Y., Zhang, H., Zhang, W., Lin, Y., Liu, S., et al.: Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. ICML 2024 (2024)
61. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., Chen, W.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. arXiv preprint arXiv:2311.16502 (2023)
62. Yue, X., Qu, X., Zhang, G., Fu, Y., Huang, W., Sun, H., Su, Y., Chen, W.: Mammoth: Building math generalist models through hybrid instruction tuning. arXiv preprint arXiv:2309.05653 (2023)
63. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023)
64. Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H.: Pointclip: Point cloud understanding by clip. In: CVPR 2022 (2022)

65. Zhang, R., Han, J., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Gao, P., Qiao, Y.: LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=d4UiXAHN2W
66. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Dong, H., Gao, P., Li, H.: Personalize segment anything model with one shot. ICLR 2024 (2023)
67. Zhang, R., Wang, L., Qiao, Y., Gao, P., Li, H.: Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. CVPR 2023 (2023)
68. Zhang, R., Wei, X., Jiang, D., Zhang, Y., Guo, Z., Tong, C., Liu, J., Zhou, A., Wei, B., Zhang, S., et al.: Mavis: Mathematical visual instruction tuning. arXiv preprint arXiv:2407.08739 (2024)
69. Zhou, A., Wang, K., Lu, Z., Shi, W., Luo, S., Qin, Z., Lu, S., Jia, A., Song, L., Zhan, M., et al.: Solving challenging math word problems using gpt-4 code interpreter with code-based self-verification. arXiv preprint arXiv:2308.07921 (2023)
70. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)