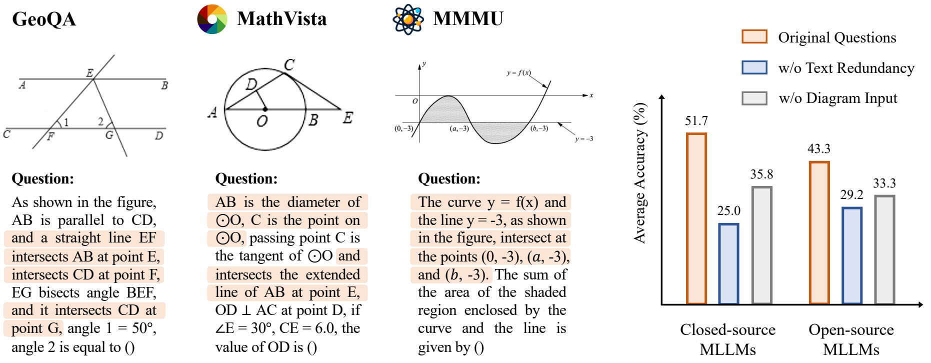
MathVerse: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems?

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

Abstract. The remarkable progress of Multi-modal Large Language Models (MLLMs) has gained unparalleled attention. However, their capabilities in visual math problem-solving remain insufficiently evaluated and understood. We investigate current benchmarks to incorporate excessive visual content within textual questions, which potentially assist MLLMs in deducing answers without truly interpreting the input diagrams. To this end, we introduce **MathVerse**, an all-around visual math benchmark designed for an equitable and in-depth evaluation of MLLMs. We meticulously collect 2,612 high-quality, multi-subject math problems with diagrams from publicly available sources. Each problem is then transformed by human annotators into six distinct versions, each offering varying degrees of information content in multi-modality, contributing to **15K** test samples in total. This approach allows MathVerse to comprehensively assess ***whether and how much MLLMs can truly understand the visual diagrams for mathematical reasoning.*** In addition, we propose a Chain-of-Thought (CoT) evaluation strategy for a fine-grained assessment of the output answers. Rather than naively judging true or false, we employ GPT-4(V) to adaptively assess each step with error analysis to derive a total score, which can reveal the inner CoT reasoning quality by MLLMs. With MathVerse, we unveil that, most existing MLLMs struggle to understand math diagrams, relying heavily on textual questions. Surprisingly, some of them even achieve 5%+ higher accuracy without the visual input. Besides, GPT-4V and MAVIS-7B achieve the best overall performance within closed-source and open-source models, respectively. We hope the MathVerse benchmark may provide unique insights to guide the future development of MLLMs. Project page: https://mathverse-cuhk.github.io.

[*] Equal contribution [‡] Project lead [†] Corresponding author



Question:

As shown in the figure, AB is parallel to CD, and a straight line EF intersects AB at point E, intersects CD at point F, EG bisects angle BEF, and it intersects CD at point G, angle 1 = 50°, angle 2 is equal to ()

Question:

AB is the diameter of ⊙O, C is the point on ⊙O, passing point C is the tangent of ⊙O and intersects the extended line of AB at point E, OD ⊥ AC at point D, if ∠E = 30°, CE = 6.0, the value of OD is ()

Question:

The curve y = f(x) and the line y = -3, as shown in the figure, intersect at the points (0, -3), (*a*, -3), and (*b*, -3). The sum of the area of the shaded region enclosed by the curve and the line is given by ()

(a) **Text Redundancy** within Existing Benchmarks (b) Ablation Study

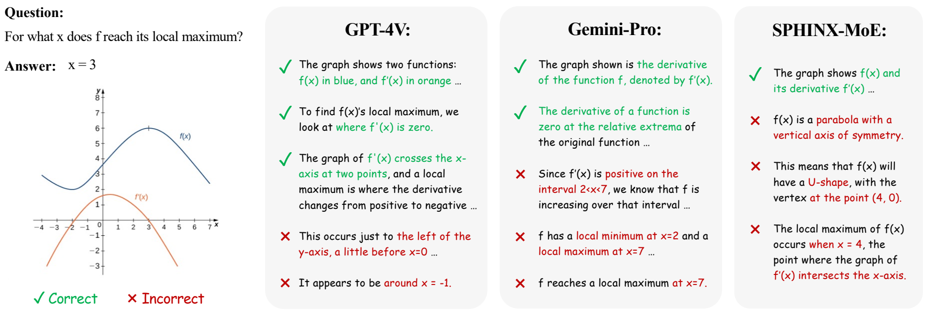
Fig. 1: (a) We showcase three examples of **Text Redundancy** (highlighted in red) within existing visual math benchmarks [6, 28, 44]. **(b)** We report an ablation study by respectively removing the redundant texts and input diagrams on 120 randomly selected problems, for closed-sourced [1, 16, 32] and open-sourced [10, 15, 25] MLLMs.

1 Introduction

With the substantial advances of big data and computational power, Large Language Models (LLMs) [2, 9, 20, 37, 38], such as ChatGPT [30] and GPT-4 [31], have emerged as a central point of interest in both industry and academia. To broaden their applicability across diverse contexts, Multi-modal Large Language Models (MLLMs) [14, 18, 40, 46] have recently become a fast-evolving track, exemplified by the latest GPT-4V [32], Gemini [16], and the open-source LLaVA [21, 22, 26] and SPHINX [15, 23]. Concurrently, a diverse array of evaluation benchmarks [11, 12, 17, 27, 36, 42] are curated to assess their visual comprehension performance across different domains. Notably, the capability to solve mathematical problems involving diagrams serves as a critical measure, offering insights into the multi-modal logical thinking prowess of MLLMs. This task demands MLLMs to accurately decode the visual elements, and correlate them with the textual condition for mathematical reasoning. Previous efforts [29, 34], e.g., GeoQA [4, 6] and UniGeo [5], concentrate on geometric problems, while the recent MathVista [28] and MMMU [44] expand the scope to encompass broader disciplines, including functions, charts, and scientific problems.

However, through our comprehensive observation and analysis, we identify three primary issues in current mathematical benchmarks for evaluating MLLMs:

i. **Do MLLMs truly see the math diagrams in evaluation?** This is the most fundamental question concerning the accurate assessment of visual math problem-solving. In Figure 1 (a), we showcase three examples from current benchmarks. We observe their texts contain too much duplicate information (highlighted in red) that is also depicted in the diagram. This redundancy might inadvertently provide MLLMs with a shortcut to resolve



Question:

For what x does f reach its local maximum?

Answer: x = 3

GPT-4V:

✓ The graph shows two functions: f(x) in blue, and f'(x) in orange ...

✓ To find f(x)'s local maximum, we look at where f '(x) is zero.

✓ The graph of f '(x) crosses the x-axis at two points, and a local maximum is where the derivative changes from positive to negative ...

× This occurs just to the left of the y-axis, a little before x=0 ...

× It appears to be around x = -1.

Gemini-Pro:

✓ The graph shown is the derivative of the function f, denoted by f'(x).

✓ The derivative of a function is zero at the relative extrema of the original function ...

× Since f'(x) is positive on the interval 2<x<7, we know that f is increasing over that interval ...

× f has a local minimum at x=2 and a local maximum at x=7 ...

× f reaches a local maximum at x=7.

SPHINX-MoE:

✓ The graph shows f(x) and its derivative f'(x) ...

× f(x) is a parabola with a vertical axis of symmetry.

× This means that f(x) will have a U-shape, with the vertex at the point (4, 0).

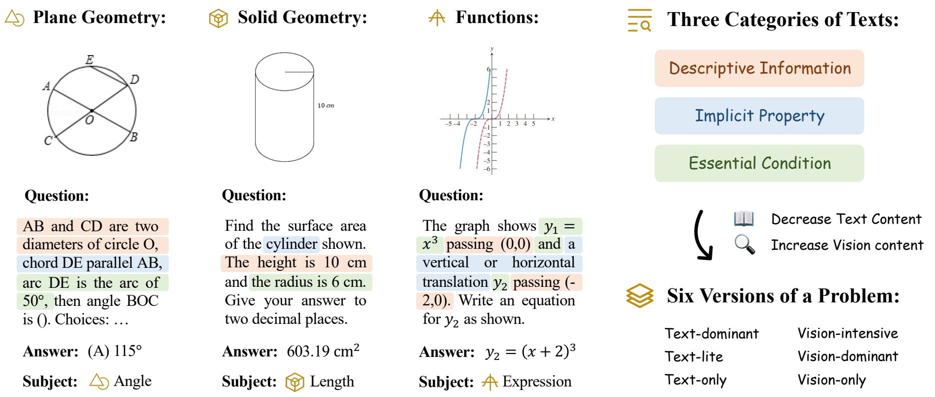
× The local maximum of f(x) occurs when x = 4, the point where the graph of f'(x) intersects the x-axis.

Fig. 2: Comparison of Visual Mathematical Reasoning by Three MLLMs. Despite the incorrect final answer, GPT-4V [32], Gemini-Pro [16], and SPHINX-MoE [15] exhibit different levels of quality in the intermediate reasoning process.

the problem by mostly reading the text, rather than interpreting the diagram. Our hypothesis gains support from the experiment in Figure 1 (b). For 40 randomly sampled problems from each benchmark, we remove such redundant texts from the question, challenging MLLMs to capture the corresponding information exclusively from visual inputs. The results reveal a significant drop in accuracy among most MLLMs (the blue column), even falling below the scores without taking diagrams as input (the grey column). This outcome suggests that ***MLLMs primarily depend on textual cues rather than the visual diagrams themselves to solve these problems in evaluation.*** Given this, we demonstrate that current visual math benchmarks might not be comprehensive enough to assess the genuine multi-modal mathematical reasoning capabilities of MLLMs.

ii. **Is it equitable to assess solely by the final answer?** Most existing multi-modal benchmarks directly compare model outputs with ground truths to derive a binary evaluation result. While this approach may suffice for general visual contexts, it falls short in math problems that require intricate step-by-step reasoning. In Figure 2, we examine three model outputs. Although they all arrive at incorrect answers in the end, they demonstrate varying levels of precision in the intermediate reasoning processes. Merely categorizing these outputs as 'Incorrect' fails to capture the nuanced differences in the reasoning quality of different MLLMs.

iii. **Do they specialize in mathematical reasoning evaluation?** GeoQA, UniGeo, and other previous works narrowly target specific aspects of plane geometry. This limits the evaluation of broader mathematical capabilities, e.g., functions and solid geometry. Instead, MathVista expands its scope by including a wide array of peripheral tasks (19 out of 28), encompassing natural images, statistic plots, and charts, which do not directly evaluate professional math skills. Furthermore, the math problems in MMMU are of college-level complexity with extensive domain-specific knowledge, potentially hindering MLLMs from fully demonstrating their reasoning capacity.



Question: AB and CD are two diameters of circle O, chord DE parallel AB, arc DE is the arc of 50°, then angle BOC is (). Choices: …

Answer: (A) 115°

Subject: Angle

Question: Find the surface area of the cylinder shown. The height is 10 cm and the radius is 6 cm. Give your answer to two decimal places.

Answer: 603.19 cm^2

Subject: Length

Question: The graph shows $y_1 = x^3$ passing (0,0) and a vertical or horizontal translation y_2 passing (-2,0). Write an equation for y_2 as shown.

Answer: $y_2 = (x+2)^3$

Subject: Expression

Fig. 3: Three Categories of Question Texts in MathVerse. According to the significance for problem-solving, we categorize the question texts into three categories, and transform each problem into six versions for evaluation, with varying content in multi-modality. We present three examples in MathVerse for illustration.

Therefore, in light of the issues discussed, we present **MathVerse**, a holistic and specialized visual math benchmark crafted to evaluate the multi-modal mathematical reasoning skills of MLLMs. This benchmark encompasses a meticulously collected dataset of 2,612 visual math problems, with 1,236 newly acquired from public question repositories and 1,376 selected from existing benchmarks, ensuring a diverse range of challenges. To specialize in mathematical reasoning, MathVerse spans three primary areas: plane geometry, solid geometry, and functions. Each problem has been rigorously reviewed by expert annotators and classified into twelve detailed categories, emphasizing different fine-grained problem-solving capabilities. Notably, MathVerse distinguishes itself by introducing two novel strategies for evaluating MLLMs.

First, we investigate the influence of textual redundancy and validate whether MLLMs can interpret the diagrams for mathematical reasoning. As illustrated in Figure 3 (Left), we categorize the textual content within the questions into three different types: *Descriptive Information*, *Implicit Property*, and *Essential Condition*. These categories, arranged in ascending order of significance for problem-solving, correspond to information directly observable from the diagram, implicit spatial properties that demand advanced visual perception, and specific measurements crucial for computing the solution, respectively. Based on this problem formulation, expert annotators progressively remove the textual information from the questions in MathVerse, while incrementally incorporating elements into the visual diagrams to ensure problems are adequately defined. As shown in Figure 3 (Right), this process results in six unique versions of each problem characterized by a reduction in textual content and an enhancement in visual elements, creating a total of 15K test samples. These delicately curated problems can indicate the various multi-modal capabilities of MLLMs, such as geometric element understanding, function curve perception, and numerical value recognition, which thoroughly unveils whether and how much they comprehend the visual diagram for mathematical reasoning.

Second, to rigorously assess the visual Chain-of-Thought (CoT) capabilities [39], we propose a **CoT Evaluation strategy** for the step-by-step reasoning assessment of MLLMs. For each model's output, we leverage GPT-4 to first extract several crucial steps exclusively from the solving process, deliberately omitting the input of the question and answer. This approach aims to mitigate the bias towards GPT-4's inherent question-answering propensities. Then, the corresponding question, diagram, and ground-truth answer are fed into GPT-4 to evaluate each identified critical step, and provide detailed error analysis. Finally, the overall score is obtained by considering every single step within reasoning. Note that, we do not pre-define a ground-truth key-step template, since each math problem may encompass a variety of solution pathways, and different MLLMs tend to exhibit variable reasoning lengths. With CoT scoring, MathVerse showcases a fine-grained evaluation of the intermediate logical deduction skills of MLLMs, demonstrating their visual mathematical CoT capabilities.

We conduct extensive experiments on MathVerse with popular closed-source [1, 16, 32] and open-source [10, 15, 25, 47] MLLMs. Comparing different problem versions, we unveil that, most existing MLLMs struggle to understand math diagrams, relying heavily on textual questions. Therein, GPT-4V [32] and MAVIS-7B [47] achieve the best overall performance within closed-source and open-source models. Surprisingly, some of the MLLMs even attain much higher results without the diagram input. With the fine-grained error analysis produced by our CoT evaluation strategy, we demonstrate such results are due to their deficient visual encoding capacity for mathematical diagrams, which instead acts as a distraction for problem-solving. In contrast, GPT-4V and InternLM-XComposer2 [10] demonstrate relatively better comprehension of the visual content for mathematical reasoning. Our experimental results suggest that inadequate mathematical visual interpretation capabilities represent the most significant impediment for MLLMs in addressing multi-modal math problems, indicating substantial potential for advancement.

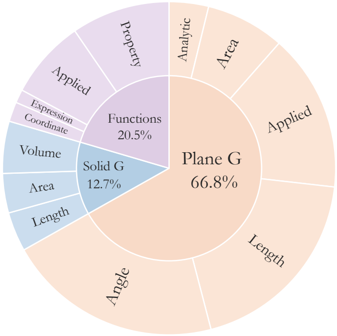
The contributions of this paper are summarized as follows:

- We investigate primary issues within existing benchmarks and introduce MathVerse, an all-around multi-modal benchmark evaluating the visual mathematical reasoning of MLLMs. The meticulously curated dataset contains 20K test problems with diagrams for a comprehensive assessment.
- By modifying problems with varying information content in multi-modality, we explore whether and how much MLLMs can understand the visual diagrams for mathematical reasoning, rather than relying on question texts.
- We propose a CoT Evaluation strategy with GPT-4 to extract and assess each key step in the reasoning process of MLLMs, which provides a fine-grained evaluation of their multi-modal mathematical CoT capabilities.

Table 1: MathVerse Statistics.

Statistic	**Number**
Total questions	2,612
- 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%)**
Total test samples	**15,672**
- **Newly annotated samples**	**10,448 (66.7%)**
- 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%)
Maximum question length	1,311
Maximum answer length	102
Average question length	204.8
Average answer length	6.3

Fig. 4: Subject Distribution of MathVerse. Solid G: Solid Geometry, Plane G: Plane Geometry.



2 MathVerse

In Section 2.1, we first present an overview of the curated visual math dataset in MathVerse. Then, in Section 2.2, we introduce our data formulation approach for investigating the visual mathematical comprehension of Multi-modal Large Language Models (MLLMs). Finally, in Section 2.3, we elaborate on the methodology of our proposed Chain-of-Thought (CoT) evaluation strategy.

2.1 Visual Math Dataset

To thoroughly assess visual mathematical proficiency, we compile a comprehensive problem set covering a broad spectrum of math subjects, diagram patterns, and specialized knowledge domains. This widespread collection for MathVerse aims to pose diverse challenges to MLLMs, ensuring a robust evaluation of their capabilities in visual contexts.

Data Composition and Categorization. MathVerse comprises a total of 2,612 visual math problems, which contribute to the final created 15K test samples. Detailed statistics for data composition are presented in Table 1. This meticulously collected dataset covers three fundamental math subjects, i.e., plane geometry (1,746), solid geometry (332), and functions (534), where the latter two are all composed of newly collected problems. The choice of these three subjects is not only due to their rigorous demands on multi-modal reasoning, but also for two other considerations. For one thing, as we specialize MathVerse in mathematical problem-solving, other peripheral tasks in MathVista [28] are not included, e.g., statistical reasoning, table question-answering, and puzzle tests. For another, we expect the evaluation can fully display the reasoning capabilities of MLLMs with moderate-level mathematical knowledge. This avoids limiting their performance with overly complex domain-specific theorems or prior commonsense knowledge. Therefore, we deliberately focus the collected problems on the high school level, excluding advanced college-level disciplines like calculus and graph theory featured in MMMU [44]. Furthermore, expert annotators subdivide the problems into twelve fine-grained categories, as depicted in Figure 4, showcasing various dimensions of visual mathematical skills.

Data Collection and Review Process. Our collection procedure for high-quality visual math problems involves a rigorous selection from both pre-existing datasets and public question repositories. In the domain of plane geometry, we initially select 750 problems from GeoQA [6], 119 from GEOS [34], and 507 from Geometry3K [29], based on their original data quality and distribution. We exclude questions that are extremely simple or excessively complex, as well as those that appear dubious or lack necessary conditions. To enhance the diversity of question types and diagram styles, we further enrich our dataset with additional 370 plane geometry problems by manually collecting from other sources[1,2,3]. Given the scarcity of solid geometry and function-related problems in existing benchmarks, we purposefully gather these two types of problems (332 and 534, respectively) from new sources[1,2,3] to address this gap. Problems that include multiple diagrams or require visual illustrations within solutions are excluded, considering the current limitations of MLLMs in resolving such information. Note that, all the newly collected problems (1,236) accompany detailed explanations. After the preliminary collection, we undertake a comprehensive review to verify the accuracy of the answers, ensure consistency between questions and diagrams, and confirm the relevance of each problem to the defined twelve categories. This meticulous review guarantees the dataset's quality and precision.

2.2 Whether MLLMs Truly See the Diagrams?

In this section, we detail our data formulation approach to transform each problem in MathVerse into six different versions with varying information content in multi-modality. In this way, we specifically explore the visual diagram understanding capabilities of MLLMs for mathematical reasoning.

Three Types of Textual Information. Considering the textual redundancy in original math problems, we first define three distinct categories for the textual information within the questions, as illustrated in Figure 3 and the following:

- **Descriptive Information** (**DI**) refers to the directly observable and clearly portrayed content in the diagram. It depicts the basic figure composition, spatial arrangement, and annotated entities, such as *the presence of geometric shapes or intersection points of functions.* Nevertheless, such information is repetitive to the visual components present in the diagram, thus regarded as redundant information for problem-solving. More importantly, it may assist MLLMs in bypassing the process of diagram interpretation, thereby undermining the assessment for visual mathematical reasoning in existing benchmarks, as evidenced in Figure 1.

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

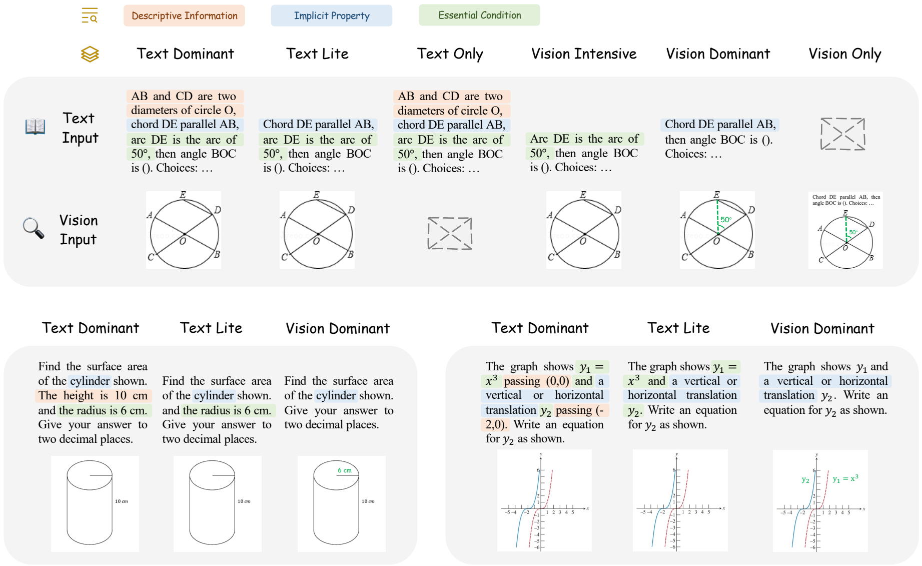

Descriptive Information
Implicit Property
Essential Condition
Text Dominant
Text Lite
Text Only
Vision Intensive
Vision Dominant
Vision Only
Text Input
AB and CD are two diameters of circle O, chord DE parallel AB, arc DE is the arc of 50°, then angle BOC is (). Choices: …
Chord DE parallel AB, arc DE is the arc of 50°, then angle BOC is (). Choices: …
AB and CD are two diameters of circle O, chord DE parallel AB, arc DE is the arc of 50°, then angle BOC is (). Choices: …
Arc DE is the arc of 50°, then angle BOC is (). Choices: …
Chord DE parallel AB, then angle BOC is (). Choices: …
Vision Input
E
D
A
O
C
B
50°
Chord DE parallel AB, then angle BOC is (). Choices: …
Text Dominant
Text Lite
Vision Dominant
Find the surface area of the cylinder shown. The height is 10 cm and the radius is 6 cm. Give your answer to two decimal places.
Find the surface area of the cylinder shown. and the radius is 6 cm. Give your answer to two decimal places.
Find the surface area of the cylinder shown. Give your answer to two decimal places.
10 cm
6 cm
Text Dominant
Text Lite
Vision Dominant
The graph shows $y_1 = x^3$ passing (0,0) and a vertical or horizontal translation y_2 passing (-2,0). Write an equation for y_2 as shown.
The graph shows $y_1 = x^3$ and a vertical or horizontal translation y_2. Write an equation for y_2 as shown.
The graph shows y_1 and a vertical or horizontal translation y_2. Write an equation for y_2 as shown.
y_2
$y_1 = x^3$

Fig. 5: Six Versions of Each Problem in MathVerse. Expert annotators meticulously transform each visual math problem within MathVerse into six versions. They contain different vision-language content for a holistic visual mathematical evaluation.

- **Implicit Property** (**IP**) involves the information that requires a higher level of visual perception but less mathematical knowledge to discern from the diagram. It signifies strong visual conditions for problem-solving, such as *the parallelism and perpendicularity between lines, the similarity and congruence among triangles, and the category and periodicity of functions.* They can, in theory, be fully extracted from the diagrams alone, given adequate capability for visual recognition and comprehension of MLLMs.
- **Essential Condition** (**EC**) denotes the specific numerical or algebraic measurements, which are indispensable conditions to derive the solution and cannot be derived from the visual diagram. This category encompasses precise values of angles, lengths, and function expressions, such as *an angle being 45 degrees, the length of BC being 6 units, and the functional equation* $f(x) = x^2 + 3$. Without these details in textual information, solving the visual math problem would be impossible.

Creating Six Versions of Each Problem. Based on the three categories, expert annotators systematically remove different textual information within questions, and incrementally incorporate the critical elements into diagrams. This approach can progressively reduce textual redundancy and information content, thereby increasingly compelling MLLMs to capture mathematical conditions from the visual input. As compared in Figure 5, we generate six versions of each problem in MathVerse, obtaining 15,672 test instances. With this curated problem set, we can provide a holistic evaluation of the genuine visual comprehension of MLLMs, and whether it can facilitate multi-modal mathematical reasoning. The details of each problem version are as follows:

- **Text-dominant Version** retains the entire textual content, *Descriptive Information*, *Implicit Property*, and *Essential Condition*, alongside the question statement. It may induce MLLMs to regard the text as the primary source of information, treating the diagram more as a supplementary visual aid. This serves as the baseline point for evaluation.

  📖 Text: **DI** + **IP** + **EC** + Question 🔍 Vision: Diagram (1)

- **Text-lite Version** diminishes the *Descriptive Information* from the Text-dominant version, assuming this information can be observed from the diagram. This creates a condensed text question without redundancy, and enforces MLLMs to interpret the diagram for basic information.

  📖 Text: **IP** + **EC** + Question 🔍 Vision: Diagram (2)

- **Text-only Version** directly discards the diagram input from the Text-dominant version. Comparing this to the Text-lite version helps identify where MLLMs mainly obtain the contextual visual information for problem-solving, the *Descriptive Information* or the diagram.

  📖 Text: **DI** + **IP** + **EC** + Question 🔍 Vision: $\varnothing$ (3)

- **Vision-intensive Version** further removes the *Implicit Property* from the Text-lite version. Without the strong visual condition in texts, MLLMs are challenged to intensively leverage their visual interpretation skills to gather sufficient cues for mathematical reasoning. The outcome demonstrates their proficiency in understanding mathematical relationships visually.

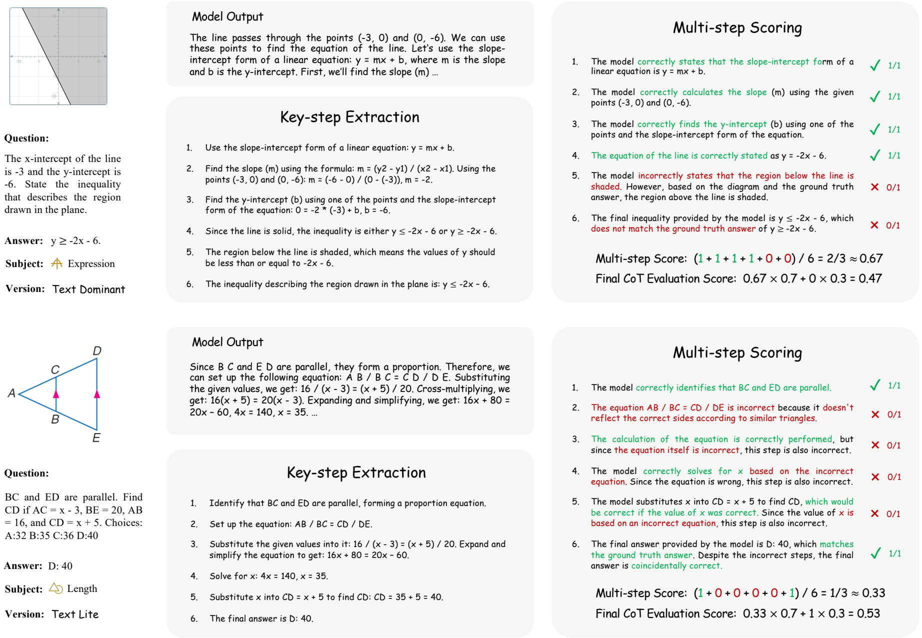
  📖 Text: **EC** + Question 🔍 Vision: Diagram (4)

- **Vision-dominant Version**, building upon the Text-lite version, excludes the *Essential Condition* from texts, instead annotating these measurements visually in diagrams. The textual content is narrowed down to *Implicit Property* and question statements. It demands MLLMs to recognize the *Essential Condition* exclusively from diagrams, and accurately correlate it with corresponding visual elements for problem-solving.

  📖 Text: **IP** + Question 🔍 Vision: Diagram + **EC** (5)

- **Vision-only Version** strips away the entire textual input, conveying the problem solely through the diagram. We render the remaining textual content in the Vision-dominant version onto the diagram, and minimize the text to an almost negligible prompt. This compels MLLMs to extract figures and recognize questions both from the visual information, standing as an ultimate test for their visual reasoning capabilities in mathematics.

  📖 Text: $\varnothing$ 🔍 Vision: Diagram + **EC** + **IP** + Question (6)



Question:

The x-intercept of the line is -3 and the y-intercept is -6. State the inequality that describes the region drawn in the plane.

Answer: y ≥ -2x - 6.

Subject: Expression

Version: Text Dominant

Model Output

The line passes through the points (-3, 0) and (0, -6). We can use these points to find the equation of the line. Let's use the slope-intercept form of a linear equation: y = mx + b, where m is the slope and b is the y-intercept. First, we'll find the slope (m) …

Key-step Extraction

1. Use the slope-intercept form of a linear equation: y = mx + b.
2. Find the slope (m) using the formula: m = (y2 - y1) / (x2 - x1). Using the points (-3, 0) and (0, -6): m = (-6 - 0) / (0 - (-3)), m = -2.
3. Find the y-intercept (b) using one of the points and the slope-intercept form of the equation: 0 = -2 * (-3) + b, b = -6.
4. Since the line is solid, the inequality is either y ≤ -2x - 6 or y ≥ -2x - 6.
5. The region below the line is shaded, which means the values of y should be less than or equal to -2x - 6.
6. The inequality describing the region drawn in the plane is: y ≤ -2x – 6.

Multi-step Scoring

1. The model correctly states that the slope-intercept form of a linear equation is y = mx + b. ✓ 1/1
2. The model correctly calculates the slope (m) using the given points (-3, 0) and (0, -6). ✓ 1/1
3. The model correctly finds the y-intercept (b) using one of the points and the slope-intercept form of the equation. ✓ 1/1
4. The equation of the line is correctly stated as y = -2x - 6. ✓ 1/1
5. The model incorrectly states that the region below the line is shaded. However, based on the diagram and the ground truth answer, the region above the line is shaded. ✗ 0/1
6. The final inequality provided by the model is y ≤ -2x - 6, which does not match the ground truth answer of y ≥ -2x - 6. ✗ 0/1

Multi-step Score: (1 + 1 + 1 + 1 + 0 + 0) / 6 = 2/3 ≈ 0.67

Final CoT Evaluation Score: 0.67 × 0.7 + 0 × 0.3 = 0.47

Question:

BC and ED are parallel. Find CD if AC = x - 3, BE = 20, AB = 16, and CD = x + 5. Choices: A:32 B:35 C:36 D:40

Answer: D: 40

Subject: Length

Version: Text Lite

Model Output

Since B C and E D are parallel, they form a proportion. Therefore, we can set up the following equation: A B / B C = C D / D E. Substituting the given values, we get: 16 / (x - 3) = (x + 5) / 20. Cross-multiplying, we get: 16(x + 5) = 20(x - 3). Expanding and simplifying, we get: 16x + 80 = 20x – 60, 4x = 140, x = 35. …

Key-step Extraction

1. Identify that BC and ED are parallel, forming a proportion equation.
2. Set up the equation: AB / BC = CD / DE.
3. Substitute the given values into it: 16 / (x - 3) = (x + 5) / 20. Expand and simplify the equation to get: 16x + 80 = 20x – 60.
4. Solve for x: 4x = 140, x = 35.
5. Substitute x into CD = x + 5 to find CD: CD = 35 + 5 = 40.
6. The final answer is D: 40.

Multi-step Scoring

1. The model correctly identifies that BC and ED are parallel. ✓ 1/1
2. The equation AB / BC = CD / DE is incorrect because it doesn't reflect the correct sides according to similar triangles. ✗ 0/1
3. The calculation of the equation is correctly performed, but since the equation itself is incorrect, this step is also incorrect. ✗ 0/1
4. The model correctly solves for x based on the incorrect equation. Since the equation is wrong, this step is also incorrect. ✗ 0/1
5. The model substitutes x into CD = x + 5 to find CD, which would be correct if the value of x was correct. Since the value of x is based on an incorrect equation, this step is also incorrect. ✗ 0/1
6. The final answer provided by the model is D: 40, which matches the ground truth answer. Despite the incorrect steps, the final answer is coincidentally correct. ✓ 1/1

Multi-step Score: (1 + 0 + 0 + 0 + 0 + 1) / 6 = 1/3 ≈ 0.33

Final CoT Evaluation Score: 0.33 × 0.7 + 1 × 0.3 = 0.53

Fig. 6: Examples of the CoT Evaluation Strategy for MathVerse. We present two outputs from Qwen-VL-Max [1] with our CoT evaluation strategy, which assesses the fine-grained reasoning capabilities with a detailed explanation for error analysis.

2.3 CoT Evaluation Strategy

Compared to visual question-answering in general scenarios, the solving process of MLLMs for mathematical problems requires nuanced, step-by-step CoT reasoning. Considering two cases in Figure 6, one arrives at the correct solution albeit through incorrect intermediary steps, while the other demonstrates the opposite phenomenon. Therefore, the binary 'Correct' or 'Incorrect' evaluative approach of existing benchmarks is inadequate to accurately examine the depth and precision of the multi-step reasoning of MLLMs. To this end, we propose a CoT Evaluation strategy to thoroughly assess their mathematical CoT skills in visual contexts, which involves two prompting phases with GPT-4(V) [31, 32].

Key-step Extraction. Given the output of an MLLM, we first employ GPT-4, the language-only version, to extract N pivotal steps within the reasoning sequence, denoted as $[s_1, s_2, \ldots, s_N]$, including the final answer s_A. Such key steps include significant computational outcomes, the identification of visual components, and critical immediate inferences. Note that, we only prompt GPT-4 with the MLLM's output, deliberately omitting the original questions, diagrams, and ground-truth answers. This approach aims to mitigate the inherent bias of GPT-4 itself towards problem-solving and visual diagram interpretation, thereby concentrating solely on the logical coherence of the model output. In addition, we do not pre-define a ground-truth key-step template for each problem, but perform the extraction adaptively for the unique output of every MLLM. Since the problem potentially encompasses diverse possible solution pathways, and different MLLMs exhibit varying reasoning lengths and styles, the rigid template would harm the CoT Evaluation accuracy.

Multi-step Scoring. After the extraction phase, we utilize GPT-4V, the multi-modal version, to evaluate each critical step and culminate a comprehensive score. We feed the extracted key steps, the original questions, diagrams, and ground-truth answers all into GPT-4V, contributing to a holistic assessment, e.g., numerical computations, logical deductions, and visual interpretations. Therein, we observe that GPT-4V occasionally struggles with accurately recognizing elements within functional diagrams, leading to unstable evaluation for related problems. We thereby annotate additional information for function problems and together feed into GPT-4V, ensuring the quality of visual evaluation. Specifically, GPT-4V assesses each N intermediate step with a binary score of '1' (correct) or '0' (incorrect), and derives the overall score by aggregating the correctness of the final answer. We formulate the scoring process as

$$\mathrm{Score}_{\mathrm{final}} = \alpha\Big(\frac{1}{N}\sum_{i=1}^{N}\mathrm{Score}(s_i)\Big) + (1-\alpha)\mathrm{Score}(s_A), \tag{7}$$

where α denotes a balancing factor between the intermediate steps and the final answer s_A. We set α as 0.7 by default to underscore the significance of CoT reasoning. As exemplified in Figure 6, besides the fine-grained scoring, the CoT evaluation can also provide a detailed error analysis of each step, which is valuable and instructive for the development of MLLMs in the field.

3 Experiments

In this section, we conduct a systematic evaluation of MLLMs on MathVerse. We showcase the direct accuracy results in Table 2 and compare the performance using Chain-of-Thought (CoT) Evaluation strategy in Table 3.

3.1 Experimental Setup

Division of the testmini Subset. To enable faster evaluation and model development validation, we extract a smaller subset termed *testmini* including 788 problems and 4,728 instances. We employ a random sampling strategy across different subfields, maintaining a sample size proportional to the overall dataset. ***In subsequent experiments, all results are based on the testmini subset.***

Evaluation Schemes. We examine foundation models across three distinct categories on MathVerse: (a) *Large Language Models (LLMs)* (only take textual questions as input), (b) *Closed-source MLLMs*, and (c) *Open-source MLLMs*.

Table 2: Accuracy Comparison on MathVerse's *testmini* Set. We report the 'Acc' results using naive 'True' or 'False' evaluations without the CoT strategy. The 'All' score is calculated without averaging the 'Text Only' version. The highest results for closed-source and open-source MLLMs is marked in red and blue. ***This is the main leaderboard of MathVerse, which is continuously being updated.***

Model	Base LLM	All	Text Dominant	Text Lite	Text Only	Vision Intensive	Vision Dominant	Vision Only
		Accuracy Evaluation						
Baselines								
Random Chance	-	12.4	12.4	12.4	12.4	12.4	12.4	12.4
Human	-	67.7	71.2	70.9	41.7	61.4	68.3	66.7
LLMs								
ChatGPT [33]	-	26.1	33.3	18.9	33.3	-	-	-
GPT-4 [31]	-	33.6	46.5	20.7	46.5	-	-	-
Closed-source MLLMs								
Qwen-VL-Plus [1]	-	11.8	15.7	11.1	14.5	9.0	13.0	10.0
Gemini-Pro [16]	-	23.5	26.3	23.5	27.3	23.0	22.3	22.2
Qwen-VL-Max [1]	-	25.3	30.7	26.1	28.9	24.1	24.1	21.4
GPT-4V [32]	-	39.4	54.7	41.4	48.7	34.9	34.4	31.6
Open-source MLLMs								
LLaMA-Adapter V2 [14]	LLaMA-7B [37]	5.8	7.8	6.3	3.9	6.2	4.5	4.4
ImageBind-LLM [19]	LLaMA-7B	10.0	13.2	11.6	12.9	9.8	11.8	3.5
mPLUG-Owl2 [41]	LLaMA-7B	10.3	11.6	11.4	13.8	11.1	9.4	8.0
MiniGPT-v2 [7]	LLaMA2-7B [38]	10.9	13.2	12.7	15.3	11.1	11.3	6.4
SPHINX-MoE [15]	Mixtral-8×7B [20]	15.0	22.2	16.4	18.3	14.8	12.6	9.1
G-LLaVA [13]	LLaMA2-7B	15.7	22.2	20.4	21.6	16.5	12.7	6.6
InternLM-XC2. [10]	InternLM2-7B [3]	16.5	22.3	17.0	16.5	15.7	16.4	11.0
ShareGPT4V [8]	Vicuna-13B [48]	17.4	21.8	20.6	14.6	18.6	16.2	9.7
Math-LLaVA [35]	Vicuna-13B	19.0	21.2	19.8	35.7	20.2	17.6	16.4
LLaVA-NeXT [25]	NH2. Yi 34B [43]	23.4	27.7	24.9	28.6	24.6	21.3	18.7
MAVIS-7B [47]	Mammoth2-7B [45]	27.5	41.4	29.1	38.6	27.4	24.9	14.6

All our experiments are conducted under a zero-shot setting. For 'Random Chance', we randomly select one option for multiple-choice questions and utilize empty for free-form questions. We also ask ten qualified college students to solve the problems independently to obtain the 'Human' performance.

3.2 Experimental Analysis

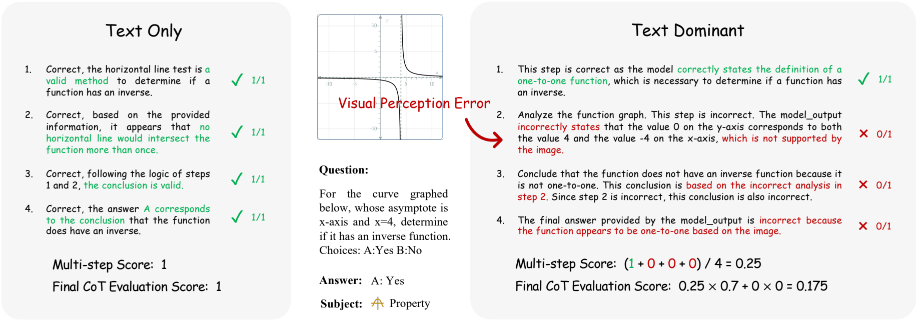
MLLMs Rely More on DI than Seeing Diagrams. Comparing the Text-lite and Text-only versions, some MLLMs encounter a larger performance drop by removing the redundant *Descriptive Information* than removing the diagram input, e.g., GPT-4V and MAVIS-7B. This pattern demonstrates that they tend to capture more visual information for mathematical reasoning from the text content, instead of seeing the diagram itself. However, for other current MLLMs, the

Table 3: CoT Evaluation on MathVerse's *testmini* Set. We employ the CoT strategy for finer-grained evaluation of MLLMs. The 'All' score is calculated without averaging the 'Text Only' version. The highest results for closed-source and open-source MLLMs are marked in red and blue.

Model	Base LLM	All	Text Dominant	Text Lite	Text Only	Vision Intensive	Vision Dominant	Vision Only
		Chain-of-Thought (CoT) Evaluation						
				LLMs				
ChatGPT [33]	-	44.9	51.3	38.5	51.3	-	-	-
GPT-4 [31]	-	52.1	63.4	40.7	63.4	-	-	-
				Closed-source MLLMs				
Qwen-VL-Plus [1]	-	21.3	26.0	21.2	25.2	18.5	19.1	21.8
Gemini-Pro [16]	-	34.5	41.6	36.8	43.7	34.0	33.0	27.4
Qwen-VL-Max [1]	-	37.5	45.4	38.9	46.1	36.7	31.9	34.6
GPT-4V [32]	-	53.6	64.6	56.7	58.9	51.5	51.2	43.9
				Open-source MLLMs				
LLaMA-Adapter V2 [14]	LLaMA-7B [37]	5.7	6.2	5.9	2.7	6.1	4.2	6.1
LLaVA-NeXT [25]	Vicuna-13B [48]	15.8	22.5	20.0	24.5	17.1	16.9	2.3
ImageBind-LLM [19]	LLaMA-7B	9.3	11.4	11.3	11.7	8.9	11.2	3.4
mPLUG-Owl2 [41]	LLaMA-7B	4.6	6.6	6.3	6.1	6.3	5.6	4.9
MiniGPT-v2 [7]	LLaMA2-7B [38]	11.0	12.1	12.0	11.7	13.1	10.3	7.4
LLaVA-1.5 [24]	Vicuna-13B	7.6	8.8	7.6	11.5	7.4	7.4	6.9
SPHINX-Plus [15]	LLaMA2-13B	12.2	13.9	11.6	14.9	11.6	13.5	10.4
G-LLaVA [13]	LLaMA2-7B	16.6	20.9	20.7	21.1	17.2	14.6	9.4
ShareGPT4V [8]	Vicuna-13B	13.1	16.2	16.2	6.6	15.5	13.8	3.7
SPHINX-MoE [15]	Mixtral-8×7B [20]	26.4	36.8	28.6	35.6	26.0	22.0	18.4
InternLM-XC2. [10]	InternLM2-7B [3]	27.4	35.8	28.2	39.0	25.3	26.4	21.3

elimination of visual input even leads to an unexpected performance improvement, e.g., Gemini-Pro and Math-LLaVA. This suggests that the unsatisfactory visual encoding for mathematical diagrams instead severely harms the original problem-solving capacity of MLLMs. As exemplified in Figure 7, from the error analysis of our CoT evaluation strategy, we observe that Gemini-Pro can deduce the correct answer exclusively by the visual information within the *Descriptive Information*. Instead, the inaccurate visual perception directly interferes with the outcome of problem-solving, turning correct answers into incorrect ones.

MLLMs are Moderately Effective at Perceiving IP. By discarding the *Implicit Property* in question texts, a negligible decline in accuracy is noted from the Text-lite to Vision-intensive versions for most MLLMs. This is because the *Implicit Property* mainly encompasses the spatial layouts and geometric relationships, which demand minimal mathematical domain knowledge for interpretation. This outcome underscores the favorable visual perception skills of MLLMs for non-mathematical elements, which is not the primary obstacle hindering MLLMs in solving visual math problems.



Text Only

1. Correct, the horizontal line test is a valid method to determine if a function has an inverse. ✓ 1/1
2. Correct, based on the provided information, it appears that no horizontal line would intersect the function more than once. ✓ 1/1
3. Correct, following the logic of steps 1 and 2, the conclusion is valid. ✓ 1/1
4. Correct, the answer A corresponds to the conclusion that the function does have an inverse. ✓ 1/1

Multi-step Score: 1

Final CoT Evaluation Score: 1

Question:

For the curve graphed below, whose asymptote is x-axis and x=4, determine if it has an inverse function. Choices: A:Yes B:No

Answer: A: Yes

Subject: Property

Text Dominant

1. This step is correct as the model correctly states the definition of a one-to-one function, which is necessary to determine if a function has an inverse. ✓ 1/1
2. Analyze the function graph. This step is incorrect. The model_output incorrectly states that the value 0 on the y-axis corresponds to both the value 4 and the value -4 on the x-axis, which is not supported by the image. ✗ 0/1
3. Conclude that the function does not have an inverse function because it is not one-to-one. This conclusion is based on the incorrect analysis in step 2. Since step 2 is incorrect, this conclusion is also incorrect. ✗ 0/1
4. The final answer provided by the model_output is incorrect because the function appears to be one-to-one based on the image. ✗ 0/1

Multi-step Score: $(1 + 0 + 0 + 0) / 4 = 0.25$

Final CoT Evaluation Score: $0.25 \times 0.7 + 0 \times 0 = 0.175$

Fig. 7: A Typical Visual Perception Error by our CoT Evaluation Strategy. The example is an output from Gemini-Pro [16], where the correct reasoning of the Text-only version is distracted by the visual perception error within the diagram.

MLLMs are Challenged to interpret EC from Diagrams. Incorporating the *Essential Condition* within diagrams challenges MLLMs to accurately identify and understand these conditions in vision modality for mathematical problem-solving. Evidence from the Vision-dominant results indicates a notable decline in the performance of most MLLMs compared to the Text-lite accuracy. This reveals their inaccurate identification of mathematical symbols and an insufficient grasp of domain-specific knowledge required to associate identified measurements with relevant concepts.

MLLMs struggle to Solve Problems Entirely by Diagrams. The scenario of Vision-only problems aligns more closely with real-world applications, where capturing an image is often more convenient than transcribing the problem into text. However, by rendering the whole question within the diagram, the mathematical problem-solving capacity of MLLMs is further diminished. This experiment unveils the great challenge for MLLMs to simultaneously understand mathematical conditions, questions, and figures from the visual input alone.

4 Conclusion

In this paper, we propose MathVerse for the visual mathematical problem-solving capacity of MLLMs. We meticulously collect high-quality math problems with diagrams spanning three primary subjects and twelve subfields. Given the issues within current benchmarks, we transform each problem into six versions, investigating how much MLLMs can interpret the visual math diagrams, and propose a CoT evaluation strategy for finer-grained reasoning assessment. By evaluating various closed-source and open-source models, MathVerse unveils that most existing MLLMs struggle to accurately understand mathematical diagrams, and even attain higher results without visual input. This indicates the potential of developing more advanced math-specific vision encoders for stronger multi-modal mathematical reasoning.

Acknowledgement

This project is funded in part by the National Key R&D Program of China Project 2022ZD0161100, by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK, by Smart Traffic Fund PSRI/76/2311/PR, and by RGC General Research Fund Project 14204021. Hongsheng Li is a PI of CPII under the InnoHK.

References

1. 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)
2. 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)
3. Cai, Z., Cao, M., Chen, H., Chen, K., Chen, K., Chen, X., Chen, X., Chen, Z., Chen, Z., Chu, P., et al.: Internlm2 technical report. arXiv preprint arXiv:2403.17297 (2024)
4. 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)
5. 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)
6. 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)
7. 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 multitask learning. arXiv preprint arXiv:2310.09478 (2023)
8. 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`
9. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. `https://lmsys.org/blog/2023-03-30-vicuna/` (March 2023)
10. 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)
11. 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)

12. Fu, C., Zhang, R., Lin, H., Wang, Z., Gao, T., Luo, Y., Huang, Y., Zhang, Z., Qiu, L., Ye, G., et al.: A challenger to gpt-4v? early explorations of gemini in visual expertise. arXiv preprint arXiv:2312.12436 (2023)
13. 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)
14. 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)
15. 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)
16. Gemini Team, G.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
17. 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/`
18. 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)
19. 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)
20. 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)
21. Li, B., Zhang, K., Zhang, H., Guo, D., Zhang, R., Li, F., Zhang, Y., Liu, Z., Li, C.: Llava-next: Stronger llms supercharge multimodal capabilities in the wild. https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/ (2024)
22. 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)
23. 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)
24. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)
25. 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/`
26. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
27. 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)
28. 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)

29. 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)
30. OpenAI: Chatgpt. `https://chat.openai.com` (2023)
31. OpenAI: Gpt-4 technical report. ArXiv **abs/2303.08774** (2023)
32. OpenAI: GPT-4V(ision) system card (2023), `https://openai.com/research/gpt-4v-system-card`
33. 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)
34. 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)
35. Shi, W., Hu, Z., Bin, Y., Liu, J., Yang, Y., Ng, S.K., Bing, L., Lee, R.K.W.: Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. arXiv preprint arXiv:2406.17294 (2024)
36. Sun, K., Pan, J., Ge, Y., Li, H., Duan, H., Wu, X., Zhang, R., Zhou, A., Qin, Z., Wang, Y., et al.: Journeydb: A benchmark for generative image understanding. Advances in Neural Information Processing Systems **36** (2024)
37. 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)
38. 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)
39. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022)
40. 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)
41. 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)
42. 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)
43. Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Li, H., Zhu, J., Chen, J., Chang, J., et al.: Yi: Open foundation models by 01. ai. arXiv preprint arXiv:2403.04652 (2024)
44. 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)
45. Yue, X., Zheng, T., Zhang, G., Chen, W.: Mammoth2: Scaling instructions from the web. arXiv preprint arXiv:2405.03548 (2024)

46. 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`
47. 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)
48. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems **36** (2024)