

MathAll: A Real-World Benchmark for Mathematical Reasoning and Cross-Modal Understanding Evaluation in Omni-MLLMs

Zhilin Lin^{1*} Zhihui Zhang^{2*} Shiliang Sun^{1†} Jing Zhao² Hao Yang³

¹State Key Laboratory of Submarine Geoscience, School of Automation and Intelligent Sensing, Shanghai Jiao Tong University

²School of Computer Science and Technology, East China Normal University

³2012 Labs, Huawei Technologies Co., Ltd., China

linz1123@sjtu.edu.cn (Zhilin Lin), shiliangsun@gmail.com (Shiliang Sun)

Abstract

Omni-MLLMs have attracted growing attention for their potential to emulate human-like multimodal perception and reasoning. However, existing mathematical reasoning benchmarks are confined to text-only or text-image formats, lacking a comprehensive evaluation of omnimodal capabilities. Moreover, these benchmarks predominantly rely on textbook-style images that differ significantly from the diverse and unstructured visuals encountered in real-world educational and human-computer interaction scenarios. To address these limitations, we propose **MathAll**, a real-world multimodal mathematics benchmark encompassing text, image, and audio modalities. By incorporating hand-drawn images and explanatory audio, MathAll simulates authentic human reasoning processes and enables comprehensive evaluation of both modality-complementary and modality-integrated reasoning. Extensive experiments demonstrate that current Omni-MLLMs struggle to effectively handle cross-modal inputs, especially when modalities contain imperfect or ambiguous information, and consequently lag behind human-level reasoning. As input complexity increases, models experience difficulties in visual perception and audio understanding, exposing inherent limitations in modality coordination and generalization. Based on these findings, we propose several promising directions for future research and model enhancement.

1. Introduction

Recent advancements in Large Language Models (LLMs) have enabled significant progress in handling complex tasks across various domains. However, real-world applications

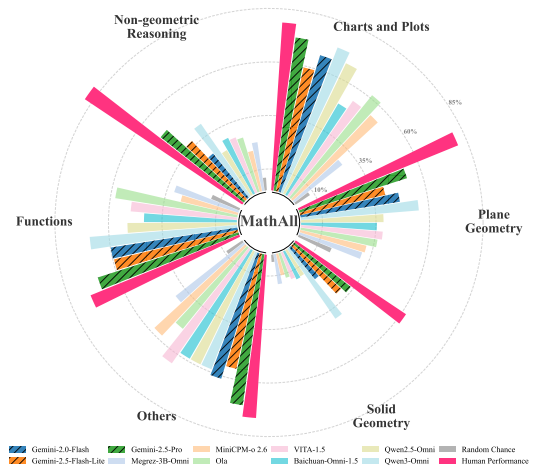


Figure 1. Accuracies of prominent Omni-MLLMs evaluated on our proposed MathAll benchmark.

increasingly require models to process and integrate multimodal inputs, such as text, images, and audio. This demand has led to the development of Multimodal Large Language Models (MLLMs), including Vision-MLLMs [16, 18, 32, 34], Audio-MLLMs [4, 25, 27], and Omni-MLLMs [14, 29, 37, 39], which aim to process and fuse information from multiple modalities within a unified framework, enhancing perception and reasoning beyond text-based understanding.

As Omni-MLLMs continue to develop, evaluating their cross-modal reasoning capabilities becomes increasingly important. Mathematical problem-solving, with deterministic answers and complex logical reasoning, has become a key benchmark for assessing model performance. Despite extensive research and diverse datasets, most existing multimodal mathematical benchmarks remain limited to text-only or text-image formats, often relying on clean, textbook-style images. This raises an essential question: are current benchmarks sufficient to assess cross-modal rea-

*Equal contribution.

†Corresponding author.

soning in Omni-MLLMs? Or do they effectively capture model weaknesses, such as errors in image comprehension and text-image integration? The answer is no. Real-world mathematical problems, especially in educational settings such as online learning or collaborative classrooms, are far more complex. They often contain hand-drawn images with inconsistent qualities, introducing ambiguity in visual understanding. At the same time, effective problem-solving relies on integrating multiple information sources, including text, verbal instructions, and visual cues, which further increases the demands on cognitive and reasoning abilities.

To address these limitations, we propose **MathAll**, a real-world multimodal benchmark designed to assess mathematical reasoning across text, image, and audio modalities. Unlike traditional benchmarks, MathAll simulates real-world classroom settings, where teachers explain mathematical concepts verbally while drawing on blackboards or tablets. Each problem in MathAll is accompanied by a hand-drawn image and an explanatory audio narration, reflecting the multimodal communication. By integrating these diverse modalities, MathAll enables the evaluation of two critical aspects of multimodal reasoning: (1) *modality-complementary reasoning*, which assesses whether a model can infer missing or ambiguous information in one modality using cues from others, such as understanding geometric relationships in an image with the help of text or audio; (2) *modality-integrated reasoning*, which examines the model’s ability to perform unified reasoning across modalities, requiring simultaneous interpretation of textual instructions, visual diagrams, and spoken explanations to solve a mathematical problem. As depicted in Figure 1, there is still a gap between existing models and human performance. This highlights that MathAll offers a more realistic and comprehensive benchmark for evaluating multimodal models on complex, real-world mathematical reasoning tasks.

In summary, our key contributions are as follows:

- We introduce MathAll, the first full-modal benchmark for mathematical reasoning, featuring 1,874 text questions, hand-drawn images, and explanatory audios. The testmini split of our benchmark is publicly available at <https://huggingface.co/datasets/SilverPlasmid/MathAll>.
- We evaluated state-of-the-art closed-source and open-source models on MathAll, finding that they still lag behind humans on full-modal mathematical tasks, especially in complex spatial reasoning with hand-drawn images, which is hard to improve through CoT.
- Ablation studies show that current models lack cross-modal audio understanding, despite audio’s potential to complement hand-drawn image reasoning. Error analysis highlights visual perception, cross-modal integration, and solution planning errors, from which we identify three practical directions for model improvement.

Benchmarks	Modalities	Size	Subjects	Question Type	Language
GSM8K [5]	T	1,319	-	OE	EN
MATH [11]	T	5,000	7	OE	EN
JEEBench [1]	T	236	-	MC,OE	EN
MMLU [10]	T	948	4	MC	EN
AGIEval [46]	T	943	4	MC,FB,OE	EN,ZH
PolyMath [36]	T	500	-	OE	EN,ZH,ES,AR,...
MMMU [43]	T+I	540	8	MC,OE	EN
SciBench [35]	T+I	217	3	OE	EN
MathVista [22]	T+I	3,402	6	MC,OE	EN
MathVision [31]	T+I	3,344	16	MC,OE	EN
MathVerse [44]	T+I	2,612	3	MC,OE	EN
CMM-Math [19]	T+I	15,213	13	OE,MC,FB,J	ZH
MathAll (Ours)	T+I,T+V,T+I+V	1,874	6	MC,OE	EN,ZH

Table 1. Comparison of our MathAll with others. For modalities, T means text, I means image and V means video. For size, only mathematical problems are counted. For subjects, only math-related subjects are included. For question type, OE: open-ended, MC: multiple-choice, FB: fill-in-the-blank, J: true-or-false. For language, EN: English, ZH: Chinese. “-” indicates that the information cannot be confirmed based on existing benchmarks.

2. Related Work

2.1. Benchmarks for Mathematical Reasoning

GSM8K [5] and MATH [11] are widely used text-only benchmarks for LLM mathematical reasoning. To incorporate visual reasoning, MathVista [22] combines math and visual tasks, MathVerse [44] tests symbol interpretation via multiple problem versions, and MathVision [31] evaluates real-world competition problem-solving skills. As models continue to advance, traditional benchmarks have gradually lost their ability to effectively differentiate performance, motivating the development of more challenging problems and complex evaluation settings. OlympiadBench [9] offers 8,476 bilingual, multimodal mathematical and physics problems, while LiveAoPSBench [23] auto-updates its dataset with evolving AoPS forum posts. MATH-Perturb [12] creates perturbation datasets to assess problem-solving abilities, and CMM-Math [19] provides a large collection of Chinese multimodal problems. MV-MATH [33] extends the visual context by offering multiple images per problem. Unlike these datasets, MathAll incorporates text, audio, hand-drawn images, offering a more realistic benchmark for evaluating cross-modal reasoning. A comparison with these benchmarks is shown in Table 1.

2.2. Multimodal Large Language Models

Recent advancements in reinforcement learning [42, 45] and MoE frameworks have accelerated the development of LLMs. Like humans using multiple senses to process information, these models are evolving from text-only input to integrating visual and auditory modalities, enhancing their perception and reasoning. Representative MLLMs, including LLaVA [18], InternVL [3], Qwen-VL [2], DeepSeek-VL [21], Gemini [29], VITA-1.5 [8], Ola [20], Qwen2.5-Omni [38], and Qwen3-Omni [39], have become indispensable tools for daily decision-making and are widely dis-

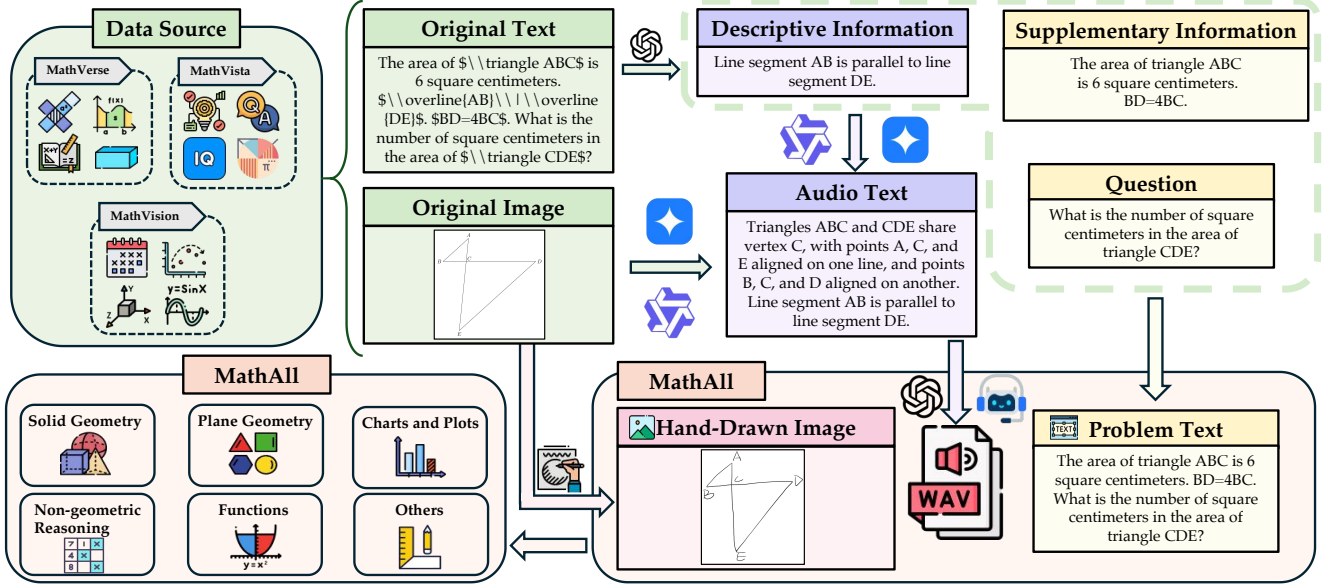


Figure 2. An overview of the design and structure of the MathAll benchmark for multimodal mathematical reasoning. The original text is segmented into descriptive information, supplementary information, and question, with the last two parts recombined to form the MathAll problem text. For the images in the original datasets, we create corresponding hand-drawn versions. Additionally, descriptive information and captioning from the original image are combined and converted into explanatory audio. Finally, each sample, encompassing all three modalities, is filtered and categorized into six mathematical subjects.

cussed and applied across domains. However, most current research [7, 13, 28, 30] primarily focuses on text-only or text-image formats, raising concerns about the reasoning capabilities, reliability, and practical applicability of these Omni-MLLMs. While Omni-SafetyBench [24] introduces the first benchmark for evaluating the safety, it still lacks an assessment of mathematical reasoning abilities.

3. The MathAll Benchmark

In this section, we introduce MathAll, the first mathematical benchmark integrating text, image, and audio modalities. The construction process is shown in Figure 2. It contains 1,874 full-modal mathematical problems across six subjects, each accompanied by a hand-drawn image and an audio clip. Thirteen distinct audio tones provide a diverse and realistic evaluation environment. Overall statistics and examples are shown in Figure 3.

3.1. Data Source

MathAll original samples are sourced from three vision-language datasets: MathVista [22], MathVerse [44], and MathVision [31]. MathVista includes math-targeted and general Visual Question Answering (VQA) problems, making it an excellent source for sample selection. MathVerse focuses on geometry and function problems, and its redundancy-removal mechanism inspires our dataset design. MathVision provides more challenging graphical

problems, validated by expert annotators, further testing reasoning ability. Since all three datasets cover a wide range of mathematical skills and are carefully curated to maintain high data quality, they are widely recognized and used as benchmarks for evaluating model performance.

3.2. Data Curation

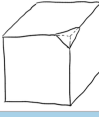
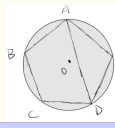
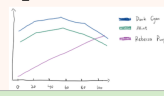
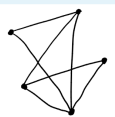
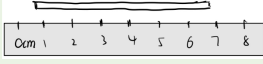
Building upon the three datasets mentioned above, we propose MathAll, an omni-mathematical benchmark that includes revised problem text, hand-drawn images, and explanatory audio. In the following paragraphs, we describe how the inputs for each modality are modified or created. Further details are provided in the Supplementary Material.

Problem Text constitutes the text modality of MathAll, containing both the question and the essential information needed for solving it. Inspired by MathVerse [44], we divide original text into *descriptive information*, *supplementary information*, and *question*. *Descriptive information* (e.g., shapes, symmetry, parallelism) can be perceived directly from images and including it in text reduces visual reasoning difficulty. *Supplementary information* (e.g., angle measures, function expressions) is not visually available but necessary for solving and must be preserved to ensure solvability. Thus, we remove *descriptive information*, retain *supplementary information*, and keep the *question* unchanged, encouraging *modality-complementary reasoning*.

Practical Approach: We prompted GPT-4o [14] to segment the original problem text. Afterwards, we verified the

Statistic	Number
Total questions	1,874
- Multiple-choice questions	983
- Free-form questions	891
- Questions in <i>testmini</i>	300
Subjects	6
Source datasets	
- MathVerse	788
- MathVista	573
- MathVision	513
Maximum question length	133
Average question length	22.7
Maximum audio text length	123
Average audio text length	31.1

(a) Key statistics of MathAll.

Solid Geometry	Plane Geometry	Charts and Plots
Question: The block initially has 8 vertices. How many vertices has the block now? Audio: 点 A 是一个木块的一个角，现在被切掉了，形成了一个三角形的面。 Question_type: free_form  Language: en Answer: 10	Question: $\angle ADE$ 的度数为 () 选择: (A) 30° (B) 32° (C) 36° (D) 40° Audio: 点 O 是正五边形 ABCDE 的中心，也是其外接圆的圆心。 Question_type: multi_choice  Language: zh Answer: C	Question: Does Rebecca Purple have the minimum area under the curve? Choices: \n(A) yes\n(B) no Audio: 点 A 的 x 轴范围从 0 到 100，包含类别包括 Mint, Rebecca Purple, 和 Dark Cyan。 Question_type: multi_choice  Language: en Answer: A
Non-geometric Reasoning	Functions	Others
Question: Each city is joined to every other city by a road. How many roads are missing? Audio: 点 A 有五座城市通过道路连接，表示为线段的集合。 Question_type: free_form  Language: en Answer: 3	Question: What is the y-intercept of this function? Audio: 点 A 的抛物线开口向上，其顶点在点 (0, 1)。 Question_type: free_form  Language: en Answer: 1	Question: Move the ruler to measure the length of the line to the nearest centimeter. The line is about () centimeters long. Audio: 点 A 的尺子显示了一个从 0 到 8 厘米的刻度。 Question_type: free_form  Language: en Answer: 6

(b) Sampled MathAll examples from each subject.

Figure 3. Detailed information about MathAll.

consistency of the question before and after segmentation using both GPT-4o and manual inspection.

Hand-Drawn Images are created by human annotators. Unlike textbook-style images, which are commonly used in most multimodal mathematical datasets, our hand-drawn images impose greater demands on models’ perceptual understanding. A distinctive feature of these images is the deliberate or natural introduction of ambiguity by human annotators. For instance, annotators are instructed to draw irregular circles (incomplete or slightly elliptical), nearly parallel lines that do not appear perfectly parallel, or perpendicular lines that are not strictly perpendicular. Introducing such ambiguity serves two purposes. First, it provides a more realistic simulation of real-world educational scenarios. Many teachers prefer hand-drawing figures on blackboards or tablets, and human learners can easily adapt to these styles. Similarly, an Omni-MLLMs are expected to be capable of understanding such imperfect visuals. Second, consistent with the rationale applied to problem text processing, this approach further evaluates the model’s *modality-complementary reasoning* ability, which requires the model to cross-reference multiple modalities when visual information alone is insufficient.

Practical Approach: All images are drawn by three human annotators, each holding a bachelor’s degree in science or engineering. They are instructed to introduce various forms of natural ambiguity. Annotators verify their drawings on two dimensions: (1) whether numerical values remain consistent with the original, and (2) whether the question’s correct answer is preserved between the original and hand-drawn versions. After independent drawing was completed, cross-checking was conducted, with each annotator reviewing the drawings of the other annotators and providing suggestions for improvement.

Explanatory Audio is designed to simulate how teachers verbally explain problems while drawing. Teachers often narrate while drawing on blackboards or tablets during teaching. For example, while sketching a plane geometry problem, a teacher might say, “This is a circle, and these two lines are parallel.” When illustrating a bar chart, they may describe it as “a bar chart showing how the female employment rate changes over time.” Such explanatory audio helps students better understand the hand-drawn images. Inspired by this, we add explanatory audio for each sample in MathAll, conveying key information intended by the image and mitigating the ambiguity introduced by hand-drawing. Importantly, the audio is designed to complement rather than replace the image. Although it plays a critical role in understanding the problem, it does not contain enough information for complete visual reconstruction. Therefore, models must integrate both the auditory and visual modalities to achieve a full understanding of the problem context. In other words, the model must reason over the correspondence between information from different modalities before proceeding to solve the question in the problem text.

Practical Approach: We employ GPT-4o-mini-tts [14] to generate explanatory audio in both English and Chinese, with varied tones to enhance naturalness and diversity. The text used for audio generation consists of two components: (1) the *descriptive information* extracted from the original problem text (see the paragraph of **Problem Text**), and (2) image captions generated from the original textbook-style images. The captions are independently generated by Qwen3-VL [40] and Gemini-2.5-Pro [6]. To minimize errors, GPT-4o is then used to extract common and consistent elements across captions. Before generating the final audio, the text undergoes multiple verification stages: content validation (checking for errors, omissions, and redundan-

Models	SG		PG		CP		NR		Fc		Ot		Avg	
	Direct	CoT	Direct	CoT	Direct	CoT	Direct	CoT	Direct	CoT	Direct	CoT	Direct	CoT
<i>Baselines</i>														
Random chance	3.5	-	16.6	-	8.0	-	6.0	-	14.2	-	9.0	-	12.4	-
Human performance*	62.0	-	80.4	-	78.6	-	88.5	-	75.0	-	76.5	-	76.8	-
<i>Closed-Source Omni-MLLMs</i>														
Gemini-2.0-Flash	19.1	35.7	46.9	56.4	66.7	79.3	26.5	33.6	59.8	64.8	61.3	62.2	46.4	56.4
Gemini-2.5-Flash-Lite	30.7	29.3	40.8	40.3	59.2	74.7	37.4	39.8	59.4	64.4	55.0	62.2	44.2	46.6
Gemini-2.5-Pro	33.6	43.8	52.7	58.8	72.3	78.4	49.4	62.7	69.0	72.6	71.2	65.8	54.9	60.9
<i>Open-Source Omni-MLLMs</i>														
Megrez-3B-Omni	14.1	13.1	30.7	25.1	28.6	20.2	22.9	19.3	31.5	24.7	40.5	39.6	28.3	23.3
Ola	12.0	18.4	36.3	39.2	61.5	62.9	26.5	26.5	57.7	48.4	48.9	58.6	37.8	40.4
MiniCPM-o 2.6	10.3	11.7	31.8	29.5	54.0	57.8	19.3	22.9	37.4	35.6	57.7	58.6	32.7	32.2
VITA-1.5	13.4	11.7	38.3	37.8	54.0	54.5	27.7	33.7	49.8	46.1	64.9	61.3	38.8	37.9
Baichuan-Omni-1.5	14.5	12.7	35.4	29.5	49.3	32.9	28.9	12.1	43.4	36.5	58.6	51.4	35.9	28.7
Qwen2.5-Omni	14.1	27.9	38.5	42.7	67.6	67.6	24.1	24.1	51.1	59.4	58.6	55.0	40.1	45.1
Qwen3-Omni	39.6	41.0	55.2	55.7	72.8	76.1	41.0	34.9	69.0	69.9	58.6	57.7	56.0	56.6

Table 2. Experimental results across different reasoning types and subjects using direct prompts and CoT prompts. SG: Solid Geometry, PG: Plane Geometry, CP: Charts and Plots, NR: Non-geometric Reasoning, Fc: Functions, Ot: Others. The highest accuracy for closed-source and open-source Omni-MLLMs is marked in red and blue respectively. * indicates evaluation on the testmini set.

cies), information deduplication, and natural-language fluency optimization through LLMs and manual review. The finalized text is then converted into speech using GPT-4o-mini-tts with randomly selected tones.

3.3. Data Filtering

Although MathVerse, MathVista, and MathVision contain high-quality mathematical problems, not all are strongly related to mathematical reasoning or common in real classroom contexts. Therefore, we filtered out problems that are irrelevant to mathematical reasoning or whose images are unlikely to appear in realistic, hand-drawn form. The remaining samples are categorized into six subjects: *Solid Geometry*, *Plane Geometry*, *Charts and Plots*, *Non-geometric Reasoning*, *Functions*, and *Others*. We manually reclassified source samples into these categories, each focusing on a distinct reasoning facet. For instance, in some *Plane Geometry* problems, models must utilize audio information to infer positional relationships between line segments; in *Charts and Plots* problems, they must interpret the topic or data represented in the chart based on audio cues. Conducting experiments grouped by subject allows us to explore model performance variations across reasoning types.

4. Experiment

4.1. Setup

Evaluation Models. For our experiment, we selected seven open-source models and three closed-source models. The

open-source models include Qwen3-Omni [39], Qwen2.5-Omni [38], Baichuan-Omni-1.5 [17], MiniCPM-o 2.6 [41], VITA-1.5 [8], Ola [20], and Megrez-3B-Omni [15]. For the closed-source models, we focused on the Gemini series [6], which is currently the only model family providing API services that support the simultaneous input of text, image, and audio. We evaluated three versions of Gemini: Gemini-2.0-Flash, Gemini-2.5-Flash-Lite, and Gemini-2.5-Pro.

Implementation Details. To comprehensively evaluate the Omni-MLLMs, we design experiments under three different settings: direct answering, Chain of Thought (CoT) answering, and 2-shot CoT answering. In the CoT setting, each model is required to independently generate responses three times in order to additionally calculate the pass@3 score. We carefully adjust the temperature and top- k parameters to ensure diversity among the generated responses. To establish a baseline for comparison, we use random chance as a reference, where we select a random option for multiple-choice questions and leave free-form questions unanswered. For benchmarking human performance, we created the testmini dataset and recruited six master’s students specializing in science or engineering to independently complete the questions. More details about the experiment are provided in the Supplementary Material.

Evaluation Metrics. Model accuracy is the primary evaluation metric. After obtaining the model outputs, we evaluate the results using both string matching and LLM-based methods to compare the predicted answers with the ground-truth labels for fairness. The string-matching implementa-

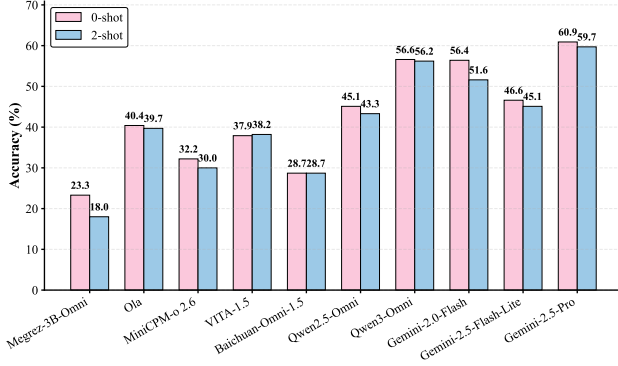


Figure 4. Performance difference between 0-shot CoT prompting and 2-shot CoT prompting.

tion is adapted from DeepSeek-Math [26], while GPT-4o is used for the LLM-based evaluation. Empirical results demonstrate that this combination ensures robust performance and minimizes misjudgments.

4.2. Main Results

Our experimental results on MathAll show that despite significant progress in full-modal multimodal reasoning models, their cross-modal reasoning capabilities remain limited. In particular, when processing hand-drawn images and audio, their performance still lag behind that of humans. Although audio was originally expected to help resolve ambiguities or noise in visual information, current models struggle to integrate audio effectively with other modalities and may even produce hallucinations during reasoning. Across all subjects, models perform poorly on tasks that demand complex perception and high-level spatial reasoning which are abilities that are critical for humans. Furthermore, we observe that Omni-MLLMs, especially open-source models, exhibit a relatively low level of instruction-following ability, which limits the benefits they derive from CoT and two-shot prompting. Additional results and detailed analyses are provided in the Supplementary Material.

Cross-Modal Understanding and Reasoning Remain Challenging. Despite their ability to handle multimodal inputs, Omni-MLLMs have significant room for improvement in effectively integrating three modalities for coherent understanding. As shown in Table 2, when prompted to generate responses with step-by-step reasoning, among closed-source models, the best-performing Gemini-2.5-Pro achieves an accuracy of 60.9%, which is 79.3% of human performance. Among open-source models, Qwen3-Omni shows significant performance improvement compared with the others, but still reaches only 73.7% of human performance. This indicates a clear gap between Omni-MLLMs and humans in handling real-world mathematical reasoning. **Performance Varies across Subjects.** Model performance

differs notably among mathematical domains. Among all subjects, models perform relatively well on *Charts and Plots* and *Others*. A possible reason is that these subjects require relatively simple reasoning processes, allowing models to infer information from images through commonsense reasoning rather than extracting additional details from the audio. In contrast, models perform poorly on *Solid Geometry* and *Non-geometric Reasoning*, reaching only 32.5% and 32.8% of human performance on average. These subjects demand deeper visual understanding, and the hand-drawn images may confuse models even though humans can interpret them easily, resulting in a significant performance gap. **CoT and Few-Shot Prompting Offer Limited Benefits.** From Table 2, we find that CoT does not significantly enhance performance for most models, particularly for models with low accuracy, and in the case of Baichuan-Omni-1.5, performance even decreases by 20.1%. Meanwhile, Figure 4 shows that few-shot CoT tends to reduce performance compared with zero-shot prompting. Most Omni-MLLMs show similar performance in zero-shot and few-shot settings, while a small number of models face a significant performance decrease when given few-shot prompts. A similar phenomenon has been observed in MathVision [31]. For open-source models, this degradation may stem from their weaker instruction-following capabilities, whereas for closed-source models, it may be due to the additional comprehension burden. More detailed analyses and discussions are provided in the Supplementary Material.

4.3. Ablation Analysis

To better understand the performance of Omni-MLLMs across three modalities of reasoning, we conducted a series of ablation experiments, whose results are presented in Table 3. The model inputs were as follows: (1) text, hand-drawn images, and explanatory audio from MathAll; (2) text and hand-drawn images from MathAll; (3) images from the original datasets with text from MathAll; (4) text and images from the original datasets; and (5) text and explanatory audio from MathAll. Experimental results demonstrate the effectiveness of the construction of each modality in MathAll and indicate that MathAll serves as an important tool for identifying the specific modalities where the models exhibit weaknesses in cross-modal reasoning.

Three vs. Two Modality Combination. Comparing the first and fourth columns reveals that, except for Qwen2.5-Omni, all Omni-MLLMs achieve lower accuracy when processing trimodal inputs in MathAll compared to traditional visual-language inputs. This reflects the difficulty the model faces in reconstructing the complete problem from multimodal information. While both input types provide complete problem-solving information, our MathAll highlights models' vulnerabilities in cross-modal integration, visual perception, and audio processing. Qwen2.5-Omni focuses

Models	Text & Image & Audio	Text & Image	Text & Image	Text & Image	Text & Audio
<i>Closed-Source Omni-MLLMs</i>					
Gemini-2.0-Flash	56.4	53.6(-2.8)	60.9(+4.5)	65.4(+9.0)	27.9(-28.5)
Gemini-2.5-Flash-Lite	46.6	45.7(-0.9)	52.1(+5.5)	58.2(+11.6)	21.9(-24.7)
Gemini-2.5-Pro	60.9	61.5(+0.6)	68.2(+7.3)	75.4 (+14.5)	29.1(-31.8)
<i>Open-Source Omni-MLLMs</i>					
Megrez-3B-Omni	23.3	25.1(+1.8)	31.1(+7.8)	32.8(+9.5)	21.4(-1.9)
Ola	40.4	36.5(-3.9)	39.7(-0.7)	41.6(+1.2)	35.8(-4.6)
MiniCPM-o 2.6	32.2	35.6(+3.4)	39.5(+7.3)	44.2(+12.0)	38.4(+6.2)
VITA-1.5	37.9	34.0(-3.9)	39.0(+1.1)	39.8(+1.9)	36.3(-1.6)
Baichuan-Omni-1.5	28.7	25.9(-2.8)	30.7(+2.0)	33.8(+5.1)	27.7(-1.0)
Qwen2.5-Omni	45.1	38.0(-7.1)	43.3(-1.8)	45.1(+0.0)	42.3(-2.8)
Qwen3-Omni	56.6	42.5(-14.1)	51.0(-5.6)	56.9(+0.3)	44.3(-12.3)

Table 3. Ablation experiments across different input modalities. Text, Image, and Audio represent results obtained from our MathAll benchmark, while the others correspond to results from the original dataset.

more effectively on these cues, showing the more robust multimodal reasoning of Qwen models.

Complete Modalities vs. Image-Lack vs. Audio-Lack. By comparing the first, second, and last columns of Table 3, we can identify how models behave under different types of modality absence. A general trend is that the completeness of the modalities generally benefits performance, as observed in Ola and Qwen2.5-Omni. However, if a model handles a particular modality poorly, adding that modality may increase confusion and ultimately lead to a decrease in performance. For instance, MiniCPM-o 2.6 exhibits degraded performance when additional image or audio are introduced. For the Gemini model, adding images greatly improves performance, while audio offers little benefit, indicating it complements but does not replace images. Closed-source models are worse than open-source models at recognizing and using audio, limiting its effectiveness.

Disturbed vs. Standard Information. By comparing the second, third and fourth columns, we observe that for textbook-style images, removing descriptive text information causes an average accuracy drop of 6.9%, indicating that although this removal forces Omni-MLLMs to actively extract information from images, they are still capable of handling this to some extent. However, evidence shows that some open-source models, such as Baichuan-Omni-1.5, often exhibit hallucinations, fabricating information that does not appear in the disturbed text. When it comes to disturbed hand-drawn images, the accuracy drops by 12.1% on average, showing that these models struggle with visual ambiguity. In such cases, explanatory audio becomes essential for accurately interpreting the entire problem.

5. Error Analysis and Future Direction

In this section, we analyze the challenges models encounter when evaluated on MathAll and propose potential future di-

rections based on these observations.

5.1. Quantitative Analysis on Qwen3-Omni

To investigate the types of mistakes models make when reasoning on MathAll, we randomly selected 100 samples with incorrect generations from Qwen3-Omni, which achieves the best performance among open-source Omni-MLLMs. We categorize the errors into five types, as illustrated in Figure 5. Visual Perception Errors account for the largest proportion, at 51%, indicating that the inability to extract critical information from visual input, especially for hand-drawn images, is the primary bottleneck for model performance. The second most prevalent error type is Cross-Modal Integration Errors, which arise when models fail to utilize information from the audio modality to resolve ambiguities in the image. These errors occur when models struggle with *modality-complementary reasoning*, and an illustrative example is provided in Section 5.2.

Other error types fall under *modality-integrated reasoning* errors, including Incorrect Solution Planning, Calculation Errors, and Contradictory Reasoning Errors, collectively accounting for 23% of all errors. These errors arise when models generate incorrect solution plans or make detailed mistakes in the reasoning path. Overall, the statistics indicate that Qwen3-Omni faces substantial challenges in recognizing key information and achieving consistent understanding across different modalities.

5.2. Qualitative Error Analysis

The MathAll benchmark, which includes text, image, and audio modalities, effectively exposes Cross-Modal Integration Errors. Figure 6 presents a typical failure example where Qwen3-Omni fails to identify correspondences between information across modalities. In this example, Qwen3-Omni is asked to determine how the water level in

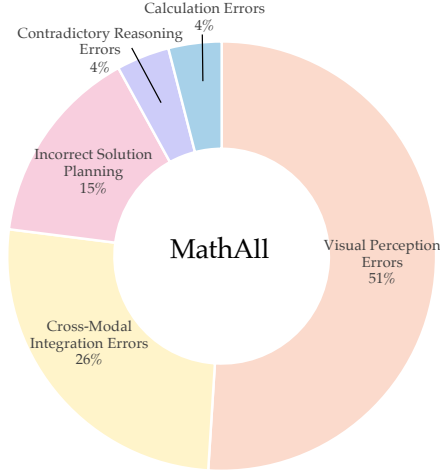


Figure 5. Distribution of Qwen3-Omni’s errors on MathAll.

a cylinder rises after pouring water from a cuboid into it. In the hand-drawn image, the shaded region represents the water to be transferred. The explanatory audio ambiguously describes the rising water level and refers to the water in the cuboid but uses the phrase *the rectangular prim*, which may mislead the model. Consequently, a model must comprehensively understand information from both the image and the audio to correctly interpret the problem. In this case, Qwen3-Omni incorrectly assumes that the cylinder is submerged in the cuboid (e.g. “the problem involves a cylinder being submerged in a rectangular prism filled with water”), failing to properly reference the conditions given in the image and audio. This type of Cross-Modal Integration Errors frequently appear in Qwen3-Omni’s generations, negatively affecting the reasoning performance of Omni-MLLMs.

5.3. Future Directions for Omni-MLLMs

Visual Perception Improvement. Visual perception remains a major challenge for Omni-MLLMs, particularly when processing hand-drawn images. Errors in extracting visual details can mislead subsequent reasoning, and current models cannot automatically detect or correct such mistakes. Training models on datasets built with our proposed method can enhance their visual understanding in noisy environments and ultimately improve their applicability in real-world multimodal scenarios.

Omni-Modal Fusion Enhancement. Omni-MLLMs can handle text, image, and audio inputs together, but they have difficulty integrating dispersed cues and maintaining consistent reasoning across modalities. Future work could develop novel multimodal fusion network architectures and explore interactive multimodal CoT methods to enhance complex multimodal understanding in real applications.

Hallucination Mitigation. Omni-MLLMs exhibit significant hallucination phenomena. Insights from our ablation

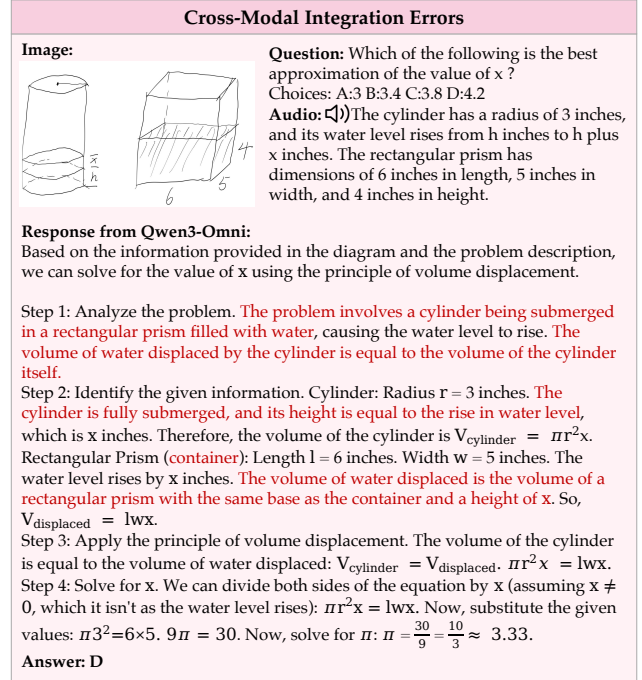


Figure 6. A typical cross-modal integration error from Qwen3-Omni on MathAll, with the error highlighted in red.

experiments show that when models find insufficient information to solve a problem, they tend to fabricate nonexistent details. This issue is more pronounced in long-chain reasoning, where models may generate information that even contradicts previously produced context. A promising direction is to assess the model’s reasoning for fidelity to the input and preceding context, leverage both long and short reasoning paths for self-check and refinement, and dynamically choose the most suitable reasoning strategy.

6. Conclusion

The introduction of MathAll marks a crucial advance in evaluating Omni-MLLMs’ mathematical reasoning, addressing a notable gap in existing benchmarks, which typically focus on limited modalities and overly idealized data. Each MathAll sample includes problem text, a hand-drawn image, and explanatory audio, allowing precise diagnosis of models’ understanding and reasoning limitations across different modalities. Our evaluation reveals that cross-modal reasoning remains a major bottleneck, particularly when multiple modalities are involved, as models often fail to leverage audio effectively. Analysis of Qwen3-Omni highlights recurring weaknesses in both modality-complementary and modality-integrated reasoning, while pointing to promising avenues for improvement. Overall, MathAll lays the groundwork for developing more robust and versatile multimodal reasoning models.

Acknowledgement

This work is supported by the National Natural Science Foundation of China under Projects 62576206 and 62476089.

References

- [1] Daman Arora, Himanshu Singh, and Mausam. Have LLMs advanced enough? A challenging problem solving benchmark for large language models. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*, pages 7527–7543, 2023. 2
- [2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 2
- [3] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 2
- [4] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023. 1
- [5] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 2
- [6] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blstein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 4, 5
- [7] Huiyu Duan, Qiang Hu, Jiarui Wang, Liu Yang, Zitong Xu, Lu Liu, Xiongkuo Min, Chunlei Cai, Tianxiao Ye, Xiaoyun Zhang, and Guangtao Zhai. FineVQ: Fine-grained user generated content video quality assessment. In *Proceedings of the conference on Computer Vision and Pattern Recognition Conference*, pages 3206–3217, 2025. 3
- [8] Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, et al. VITA-1.5: Towards GPT-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025. 2, 5
- [9] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. OlympiadBench: A challenging benchmark for promoting agi with Olympiad-level bilingual multimodal scientific problems. In *Proceedings of the annual meeting of the Association for Computational Linguistics*, pages 3828–3850, 2024. 2
- [10] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations*, 2021. 2
- [11] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021. 2
- [12] Kaixuan Huang, Jiacheng Guo, Zihao Li, Xiang Ji, Jiawei Ge, Wenzhe Li, Yingqing Guo, Tianle Cai, Hui Yuan, Runzhe Wang, et al. MATH-Perturb: Benchmarking LLMs’ math reasoning abilities against hard perturbations. In *Proceedings of the International Conference on Machine Learning*, 2025. 2
- [13] Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. Online video understanding: OVBench and VideoChat-Online. In *Proceedings of the conference on Computer Vision and Pattern Recognition*, pages 3328–3338, 2025. 3
- [14] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1, 3, 4
- [15] Boxun Li, Yadong Li, Zhiyuan Li, Congyi Liu, Weilin Liu, Guowei Niu, Zheyue Tan, Haiyang Xu, Zhuoyu Yao, Tao Yuan, et al. Megrez-omni technical report. *arXiv preprint arXiv:2502.15803*, 2025. 5
- [16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning*, pages 19730–19742, 2023. 1
- [17] Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025. 5
- [18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 2
- [19] Wentao Liu, Qianjun Pan, Yi Zhang, Zhuo Liu, Ji Wu, Jie Zhou, Aimin Zhou, Qin Chen, Bo Jiang, and Liang He. CMM-MATH: A Chinese multimodal math dataset to evaluate and enhance the mathematics reasoning of large multimodal models. *arXiv preprint arXiv:2409.02834*, 2024. 2
- [20] Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment. *arXiv preprint arXiv:2502.04328*, 2025. 2, 5
- [21] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-VL: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 2
- [22] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel

- Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *Proceedings of the International Conference on Learning Representations*, 2024. 2, 3
- [23] Sadegh Mahdavi, Muchen Li, Kaiwen Liu, Christos Thramoulidis, Leonid Sigal, and Renjie Liao. Leveraging online Olympiad-level math problems for LLMs training and contamination-resistant evaluation. In *Proceedings of the International Conference on Machine Learning*, 2025. 2
- [24] Leyi Pan, Zheyu Fu, Yunpeng Zhai, Shuchang Tao, Sheng Guan, Shiyu Huang, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Felix Henry, et al. Omni-safetybench: A benchmark for safety evaluation of audio-visual large language models. *arXiv preprint arXiv:2508.07173*, 2025. 3
- [25] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning*, pages 28492–28518, 2023. 1
- [26] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 6
- [27] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*, 2023. 1
- [28] Xiaqiang Tang, Jian Li, Keyu Hu, Nan Du, Xiaolong Li, Xi Zhang, Weigao Sun, and Sihong Xie. CogniBench: A legal-inspired framework and dataset for assessing cognitive faithfulness of large language models. In *Proceedings of the annual meeting of the Association for Computational Linguistics*, 2025. 3
- [29] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 2
- [30] Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving. *Advances in Neural Information Processing Systems*, 37:7821–7846, 2024. 3
- [31] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024. 2, 3, 6
- [32] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1
- [33] Peijie Wang, Zhong-Zhi Li, Fei Yin, Dekang Ran, and Cheng-Lin Liu. MV-math: Evaluating multimodal math reasoning in multi-visual contexts. In *Proceedings of the conference on Computer Vision and Pattern Recognition*, pages 19541–19551, 2025. 2
- [34] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024. 1
- [35] Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. SciBench: Evaluating college-level scientific problem-solving abilities of large language models. In *Proceedings of the International Conference on Machine Learning*, pages 50622–50649, 2024. 2
- [36] Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, et al. Polymath: Evaluating mathematical reasoning in multilingual contexts. *arXiv preprint arXiv:2504.18428*, 2025. 2
- [37] Zhifei Xie and Changqiao Wu. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *arXiv preprint arXiv:2410.11190*, 2024. 1
- [38] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. 2, 5
- [39] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 1, 2, 5
- [40] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 4
- [41] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A GPT-4v level MLLM on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 5
- [42] Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. 2
- [43] Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruochi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 2
- [44] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal LLM truly see the diagrams in visual math problems? In *Proceedings of the European Conference on Computer Vision*, pages 169–186, 2024. 2, 3
- [45] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025. 2
- [46] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan

Duan. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics*, pages 2299–2314, 2024. [2](#)