

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

Zhilin Lin, Zhihui Zhang, Shiliang Sun, Jing Zhao, Hao Yang
Supplementary Material

Table of Contents in Appendix

A More Details about MathAll Benchmark	2
A.1 Detailed Statistics	2
A.2 Data Collection Details	2
A.2.1. Details of Creating Tri-modal Inputs	2
A.2.2. Data Filtering Process	3
A.2.3. Construction of the Testmini Subset .	3
A.3 Introduction of Subjects	3
A.4 Comparison with Existing Datasets	4
B Additional Experimental Details	4
B.1. Experimental Setup	4
B.1.1. Models	4
B.1.2. Prompt Design	5
B.1.3. Model Hyperparameters	5
B.1.4. Evaluation Details	5
B.2. Additional Results	5
B.2.1. Results on the Testmini Subset	5
B.2.2. Additional Ablation Results	6
B.2.3. Pass@3 Results	7
B.2.4. Results on Vision-Language Models	7
B.3. More Details and Analysis	7
C Additional Analysis and Case Study	7
C.1. Representative Success Case	7
C.2. Overview of Error Types	7
C.3. Case Study of Gemini-2.5-Pro	8

A. More Details about MathAll Benchmark

A.1. Detailed Statistics

In this subsection, we present more detailed statistics of the MathAll benchmark. First, Figure 7 and 8 show the distributions of problem text lengths and audio lengths (statistics based on the number of words in the audio transcript). Since we remove descriptive information from the problem text, its length is shorter on average than in the source datasets. In particular, we note that there is one sample with a text length of only 1 (the text is “BF=”). When encountering such problem text, models face a greater challenge. As for the audio, it exhibits a wider range of lengths, reflecting varying levels of detail needed to describe the images across different problems, and thus showing different features of each subject.

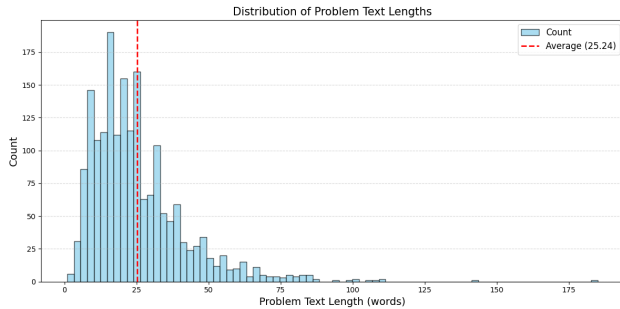


Figure 7. Distribution of problem text lengths in MathAll.

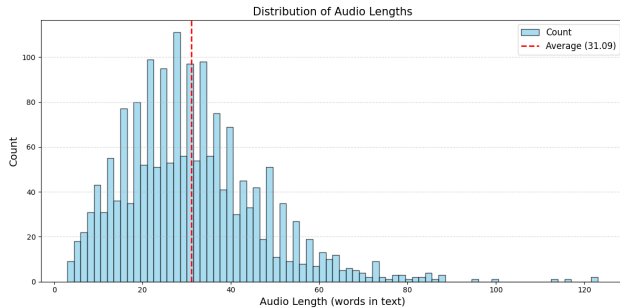


Figure 8. Distribution of audio lengths in MathAll.

Table 4 provides additional statistics for MathAll. Regarding the subject distribution, we design plane geometry problems to make up half of the dataset because they appear most frequently in real classrooms that MathAll aims to simulate, and they can more comprehensively evaluate models’ perceptual and reasoning abilities. Among the remaining subjects, solid geometry, charts and plots, and functions appear in similar proportions, while non-geometric reasoning and others are less frequent due to their lower occurrence in classroom settings. For all multiple-choice questions, we also examine the proportion of each

Statistic	Number
Total Questions	1,874
- Solid Geometry	283
- Plane Geometry	965
- Charts and Plots	213
- Non-geometric Reasoning	83
- Functions	219
- Others	111
Multiple-choice Questions	983
- Answer A	248 (25.2%)
- Answer B	275 (28.0%)
- Answer C	243 (24.7%)
- Answer D	163 (16.6%)
- Answer Other	54 (5.5%)

Table 4. More statistics of MATHALL.

option in the correct answers. In MathAll, the number of answer options varies across problems: some include three options (A/B/C), others four (A/B/C/D), and a small subset includes more options (e.g. A/B/C/D/E). Table 4 shows that even though the problems are sourced from other datasets and the original option ordering is preserved, the answer choices still exhibit a relatively uniform distribution.

A.2. Data Collection Details

A.2.1. Details of Creating Tri-modal Inputs

In this section, we will introduce the process of collecting tri-modal data in detail.

Problem Text originates from the text in the source datasets and is processed by GPT-4o, followed by verification by human annotators. For each sample, we prompt GPT-4o to split the full text into three parts: descriptive information, supplementary information, and the question. To ensure a fair comparison with the original datasets, GPT-4o is instructed to keep the same notation style in the supplementary information and question as in the original text input. To make the audio text readable, GPT-4o is required to convert mathematical notations into natural-language expressions in the descriptive information. The complete text-splitting prompt is provided in Table 10. After this procedure, the problem text in MathAll is formed by combining the supplementary information and the question. To reduce the risk of errors in the LLM-based splitting process, we prompt GPT-4o again to check the consistency between the text before and after processing, and the consistency-checking prompt is shown in Table 10. For cases that GPT-4o identifies as inconsistent, human annotators perform an additional review and refinement. This procedure ensures the correctness of the final problem text.

Hand-Drawn Images are created to simulate real-world educational scenarios. Three human annotators are instructed to draw these images based on the original im-

ages from the source datasets. During this process, they are asked to preserve as much information from the original image as possible while naturally introducing some ambiguity. To avoid unnecessary work, annotators may omit details that are less important for reasoning, such as fine axis ticks in function graphs or excessive colors in statistical charts. However, annotators must ensure that the answer to the problem remains unchanged after simplifying the image. In other words, if they omit any information in the hand-drawn image, they must verify the question to confirm that the omission does not make the problem unsolvable or change the correct answer. Each annotator is asked to self-check that key information is preserved after drawing. Subsequently, images are exchanged among annotators for cross-checking. Since relying on inter-annotator agreement scores are less suitable due to the continuous and perception-dependent nature of ambiguity, we report a cross-checking pass rate of approximately 95%, indicating high consistency in preserving answer correctness and problem solvability.

Explanatory Audio serves as an additional data source used to confirm ambiguous information in the hand-drawn images. We adopt a pipeline that first collects the audio text and then converts it to audio using GPT-4o-mini-tts. The audio text consists of two parts: descriptive information extracted from the original text, and an image caption generated by advanced LLMs. The latter is included to supplement cases in which the descriptive information does not cover all key ambiguous details in the image. To produce the image caption, we prompt Qwen3-VL and Gemini-2.5-Pro to describe the original image independently, and then prompt GPT-4o to extract the common content from the two captions to reduce potential hallucinations. These prompts are provided in Table 11. After obtaining the raw audio text, it undergoes a series of checks and refinements. The first step is content validation, where human annotators review the audio text sample by sample to identify and correct the following types of issues: errors, where the text includes false information due to LLM hallucinations; omissions, where key information is missing and the problem becomes unsolvable; and redundancies, where the audio text directly reveals the answer. The second step involves information deduplication and natural-language fluency optimization. Information deduplication focuses on removing overlapping content between the descriptive information and the image caption. Fluency optimization adjusts the text to make the audio sound more like a teacher explaining a diagram on the blackboard during class. We further apply an LLM-human consistency check, similar to the question consistency check used in processing the problem text, to ensure that the key information remains unchanged. The prompts used in this step are shown in Table 12. Once the audio text is finalized, it is converted into audio using

GPT-4o-mini-tts. For English samples, we randomly select from the tones *ash*, *ballad*, *coral*, *fable*, *onyx*, *nova*, *sage*, *shimmer*, *verse*, *marin*, and *cedar*. For Chinese samples, we randomly choose between *alloy* and *echo* since their number is relatively small.

A.2.2. Data Filtering Process

To ensure the overall quality of the dataset, we carefully filter out samples that are not directly related to mathematical reasoning or that are unlikely to appear in real classroom settings. This step helps us keep the dataset closer to what students normally see during learning. For example, when we process data from MathVista, we remove samples that use photos, cartoon characters, or complicated charts from research papers, because these styles rarely show up in everyday teaching materials. We also delete samples whose images contain too many decorative elements or look unrealistic for a school environment. Several examples of the removed items are shown in Figure 9.

A.2.3. Construction of the Testmini Subset

To facilitate analysis, we extract 300 samples from the full dataset to construct the testmini subset. Random sampling is performed across all subjects, maintaining proportional representation based on each subject’s overall distribution. In experiments shown in the main body of this paper, all model evaluations are conducted on the full dataset, while human evaluations are carried out on the testmini subset.

A.3. Introduction of Subjects

In this subsection, we provide a detailed introduction to each subject, and examples for each subject can be found in the main text of the paper.

Solid Geometry focuses on solving problems based on geometric solids such as spheres, cones, and cylinders. This subject places high demands on the model’s 3D perception capabilities, especially when dealing with hand-drawn images. Typical problems include calculating the volume or surface area of a solid figure.

Plane Geometry focuses on plane figures such as lines, circles, and combinations of shapes abstracted from real-world scenarios. This subject requires models to perform relatively complex reasoning even though the image itself demands weaker perceptual ability. Hand-drawn images may introduce ambiguous information, such as approximate shapes. Typical problems include calculating angle measures or the lengths of line segments.

Charts and Plots focuses on extracting information from statistical charts and plots, including line charts, pie charts, histograms, scatter plots, data tables, etc. Unlike Plane Geometry, this subject examines the model’s perception abilities, particularly the ability to recognize digits and words in various forms. Typical problems involve comparing category values or reading the exact value of a bar.

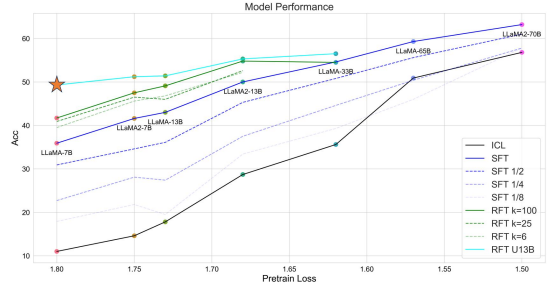


Figure 1: The key findings of scaling relationship on learning math reasoning ability with LLMs.

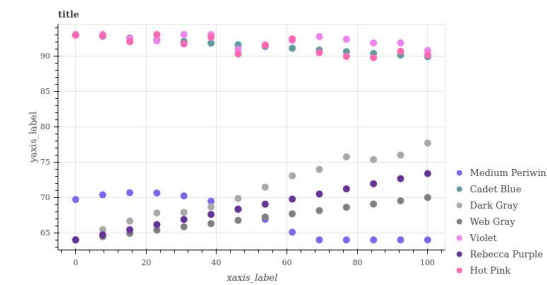


Figure 9. Some examples from MathVista that we filter out when constructing MathAll. Top Left: This example evaluates a model’s perception of realistic lighting effects, which cannot be reliably represented in hand-drawn images. Bottom Left: This sample includes cartoon characters that rarely appear in real classroom mathematics tasks. Top Right & Bottom Right: These are figures extracted from academic papers and use multiple colors and complex layouts, which are generally not hand-drawn.

Non-geometric Reasoning focuses on identifying patterns in graphical examples to solve the given problem. It requires models to form a complex perception of the images and perform visual reasoning. This subject tests a model’s ability to conduct spatial imagination based on images containing ambiguous elements. Typical problems include determining the next graph that follows a given pattern.

Functions focuses on solving problems involving function graphs. Models need to understand key properties of the function, such as the shape of the graph and its passing points. This subject relies more on processing information distributed across the image and audio modalities. Typical problems include determining monotonicity and parity, or analyzing the period and range of a function.

Others includes problems that are less directly related to mathematical reasoning but are still common in real classroom settings. This subject tests the model’s general abilities when tackling problems related to daily life or other scientific fields. Typical problems include reading time from a clock, measuring length with a ruler, or solving simple physics questions.

A.4. Comparison with Existing Datasets

In this subsection, we reiterate how MathAll differs from existing datasets to highlight its unique features.

First, unlike all existing mathematical reasoning bench-

marks, MathAll is the first to use three modalities (text, image, and audio) as input. This brings two main advantages. One is that it can comprehensively evaluate the abilities of Omni-MLLMs, since models must correctly understand all three modalities to produce the right answers. Another advantage is that this input format better reflects real-world usage. In real classrooms or human-computer interaction settings, people often use a hand-drawn image along with oral explanations to present a mathematical problem, which is the scenario MathAll aims to simulate.

Second, unlike its source datasets, i.e., MathVista, MathVerse and MathVision, MathAll significantly modifies the content in each modality. Although the raw samples are collected from these datasets, MathAll only preserves the final questions and answers. Beyond this, both the text and image inputs are redesigned, and an additional audio input is added. Therefore, MathAll maintains a high degree of originality.

B. Additional Experimental Details

B.1. Experimental Setup

B.1.1. Models

Using MathAll, we evaluate the performance of Omni-MLLMs on full-modal mathematical reasoning tasks. We

Model	Parameters	Source	URL
Qwen3-Omni	30B-A3B	local checkpoint	https://github.com/QwenLM/Qwen3-Omni
Qwen2.5-Omni	7B	local checkpoint	https://github.com/QwenLM/Qwen2.5-Omni
Baichuan-Omni-1.5	7B	local checkpoint	https://github.com/baichuan-inc/Baichuan-Omni-1.5
MiniCPM-o 2.6	8B	local checkpoint	https://github.com/OpenBMB/MiniCPM-V
VITA-1.5	8B	local checkpoint	https://github.com/VITA-MLLM/VITA
Ola	7B	local checkpoint	https://github.com/Ola-Omni/Ola
Megrez-3B-Omni	3B	local checkpoint	https://github.com/infinigence/Infini-Megrez-Omni
Gemini-2.0-Flash	-	gemini-2.0-flash	https://ai.google.dev
Gemini-2.5-Flash-Lite	-	gemini-2.5-flash-lite	https://ai.google.dev
Gemini-2.5-Pro	-	gemini-2.5-pro	https://ai.google.dev

Table 5. The source and size of the models used in the evaluation.

need the models to process text, image and audio at the same time. Due to this requirement, we collect 7 open-source Omni-MLLMs and 3 closed-source models from Gemini series for evaluations. Their details are shown in Table 5. It should be noted that even though Qwen3-Omni supports API inference in their official website, it can only receive at most two modalities as input, so we choose to download the weights and deploy it locally.

B.1.2. Prompt Design

The prompts we use for evaluation under each setting are listed in Tables 13 and 14. The two-shot prompt is constructed by first sampling two illustrative examples from external datasets, and we incorporate the image and audio descriptions in the prompt rather than the raw image and audio files to represent the input.

B.1.3. Model Hyperparameters

The hyperparameters we set in generating responses for different Omni-MLLMs are listed in Table 9. When generating based on CoT prompting, we set the temperature > 0 to ensure the generation diversity in order to calculate pass@3 score. Compared with Gemini models, open-source models have a more strict output token limitation since they tend to generate relatively shorter responses when the problems are solvable for them.

B.1.4. Evaluation Details

To ensure a fair comparison between models, accuracy must be computed consistently and appropriately. Given a JSONL file containing questions, ground truth answers, and responses generated by the model, we evaluate correctness by comparing each model’s output to the corresponding ground truth on a sample-by-sample basis. This is achieved through the following steps.

Step 1: Answer extraction. Since a model generation may include a reasoning process, we must extract the final

answer from it. We begin by applying the string-matching method from DeepSeek-Math¹. However, because models often produce answers in highly diverse formats, string matching may fail in certain cases. When this occurs, we prompt GPT-4o to extract the final answer from the model’s output.

Step 2: Answer comparison. We then check whether the extracted answer matches the ground truth. As before, we initially apply the same string-matching method as in DeepSeek-Math. Since string matching cannot handle cases where different expressions have the same meaning or when a model outputs the exact value instead of selecting an option, we again employ GPT-4o to double-check samples flagged as inconsistent by string matching. Our observations suggest that this approach effectively reduces misjudgments.

The prompts used for GPT-4o-based answer extraction and answer comparison are listed in Table 15 and 16, which are derived and modified from MathVerse official code².

B.2. Additional Results

B.2.1. Results on the Testmini Subset

To enable a fair performance comparison between Omni-MLLMs and humans, we present all models’ performance on the testmini subset in Table 6. Since the number of samples in the testmini subset is relatively small, the models’ performance carries a higher degree of uncertainty. From the table, we notice that, except for Baichuan-Omni-1.5, most models show a slight improvement in their performance, while the three main conclusions drawn from the results remain unchanged. First, cross-modal understanding and reasoning remain challenging, there is still a gap between performance between Omni-MLLMs and humans.

¹<https://github.com/deepseek-ai/DeepSeek-Math>

²<https://github.com/ZrrSkywalker/MathVerse>

Models	SG		PG		CP		NR		Fc		Ot		Avg	
	Direct	CoT	Direct	CoT	Direct	CoT	Direct	CoT	Direct	CoT	Direct	CoT	Direct	CoT
<i>Baseline</i>														
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	28.3	37.0	43.8	56.9	71.4	85.7	30.8	61.5	63.9	69.4	82.4	82.4	48.7	60.3
Gemini-2.5-Flash-Lite	37.5	31.3	39.3	38.0	48.2	85.2	33.3	66.7	61.9	71.4	76.9	76.9	45.5	51.0
Gemini-2.5-Pro	32.6	41.3	49.7	60.1	82.9	80.0	52.9	84.6	66.7	77.8	82.4	76.5	55.0	63.7
<i>Open-Source Omni-MLLMs</i>														
Megrez-3B-Omni	19.6	19.6	33.3	29.4	22.9	20.0	30.8	38.5	33.3	27.8	53.0	52.9	31.0	28.3
Ola	17.4	23.9	34.0	36.6	57.1	57.1	30.8	53.9	52.8	55.6	76.5	82.4	38.7	42.7
MiniCPM-o 2.6	15.2	15.2	29.4	24.8	60.0	68.6	30.8	46.2	38.9	38.9	70.6	76.5	34.3	34.0
VITA-1.5	13.0	10.9	34.6	34.0	60.0	65.7	46.2	30.8	52.8	50.0	88.2	61.3	40.0	38.7
Baichuan-Omni-1.5	19.6	10.9	32.0	24.2	34.3	34.3	23.1	7.7	38.9	38.9	58.8	70.6	32.3	27.0
Qwen2.5-Omni	23.9	43.5	34.6	39.9	65.7	71.4	30.8	38.5	55.6	61.1	76.5	58.8	41.3	47.7
Qwen3-Omni	41.3	47.8	53.6	57.5	77.1	80.0	69.2	61.5	75.0	80.6	76.5	70.6	59.0	62.3

Table 6. Experimental results on testmini 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.

Models	<u>T&I&A</u>	<u>T & I&A</u>	<u>T&I&A*</u>	<u>T&I</u>	<u>T& I</u>	<u>T& I</u>	<u>T&A</u>	<u>A^Δ</u>	<u>I^Δ</u>
<i>Closed-Source Omni-MLLMs</i>									
Gemini-2.0-Flash	60.3	68.3(+8.0)	56.3(-4.0)	56.3(-4.0)	62.0(+1.7)	64.0(+3.7)	31.3(-29.0)	23.0(-37.3)	57.2(-3.1)
Gemini-2.5-Flash-Lite	51.0	62.6(+11.6)	48.0(-3.0)	47.3(-3.7)	54.0(+3.0)	59.3(+8.3)	25.0(-26.0)	16.3(-34.7)	45.6(-5.4)
Gemini-2.5-Pro	63.7	77.0(+13.3)	63.6(-0.1)	61.7(-2.0)	73.0(+9.3)	71.0(+7.3)	36.4(-27.3)	20.6(-43.1)	63.3(-0.4)
<i>Open-Source Omni-MLLMs</i>									
Megrez-3B-Omni	28.3	27.0(-1.3)	24.0(-4.3)	26.3(-2.0)	29.7(+1.4)	30.7(+2.4)	26.0(-2.3)	19.3(-9.0)	11.0(-17.3)
Ola	37.0	41.7(+4.7)	37.7(+0.7)	38.0(+1.0)	39.3(+2.3)	39.7(+2.7)	36.3(-0.7)	30.7(-6.3)	33.3(-3.7)
MiniCPM-o 2.6	42.7	46.7(+4.0)	40.0(-2.7)	34.3(-8.4)	42.3(-0.4)	46.0(+3.3)	41.0(-1.7)	29.7(-13.0)	32.7(-10.0)
VITA-1.5	38.7	36.0(-2.7)	39.3(+0.6)	34.7(-4.0)	43.3(+4.6)	41.3(+2.6)	38.3(-0.4)	27.0(-11.7)	26.0(-12.7)
Baichuan-Omni-1.5	27.0	35.7(+8.7)	29.7(+2.7)	25.0(-2.0)	30.7(+3.7)	35.0(+8.0)	27.3(+0.3)	20.7(-6.3)	24.7(-2.3)
Qwen2.5-Omni	47.7	50.5(+2.8)	50.7(+3.0)	38.7(-9.0)	44.0(-3.7)	46.7(-1.0)	46.0(-1.7)	33.0(-14.7)	41.7(-6.0)
Qwen3-Omni	62.3	65.7(+3.4)	59.0(-3.3)	47.3(-15.0)	54.7(-7.6)	60.0(-2.3)	49.3(-13.0)	30.3(-32.0)	47.3(-15.0)

Table 7. Ablation results on MathAll testmini under different modality combinations. **T**, **I**, and **A** are sourced from the original datasets, while **T**, **I**, and **A** are constructed in MathAll. **A*** denotes recorded audio. **I^Δ** and **A^Δ** represent the visual-only and audio-only settings, respectively.

Second, performance varies across subjects, with problems in subjects like solid geometry and non-geometric reasoning being tougher for more models. Finally, CoT prompting provides limited benefits, particularly for open-source Omni-MLLMs.

B.2.2. Additional Ablation Results

To study the impact of synthetic versus human audio on model performance, we collect recordings from six qualified graduate students in noisy classroom-like settings (**A***, Table 7). While human audio yields more realistic performance than TTS, the performance differences across all

models are marginal, suggesting that TTS does not overestimate the contribution of audio.

For fair comparison, we apply modality transformations: converting textual information into visual-only charts (**I^Δ**), and converting question text into audio (**A^Δ**). Table 7 shows that relying solely on audio is insufficient for mathematical reasoning, as images provide essential structural and semantic cues. Moreover, extracting information from images is more challenging than from text.

Additionally, comparing original and hand-drawn images (Table 7, first two columns) shows hand-drawing gen-

erally degrades performance. While occasional minor gains likely reflect noise from testmini’s small size, overall decline suggests that hand-drawn inputs increase perceptual uncertainty for math reasoning.

B.2.3. Pass@3 Results

To more comprehensively evaluate a model’s robustness and its ability to explore diverse reasoning paths, we additionally calculate the pass@3 score. Unlike the standard pass@1 accuracy, which reflects the model’s performance in a single generation, pass@3 considers multiple sampled outputs and measures the probability that the model produces at least one correct answer within three attempts. The results are provided in Figure 10.

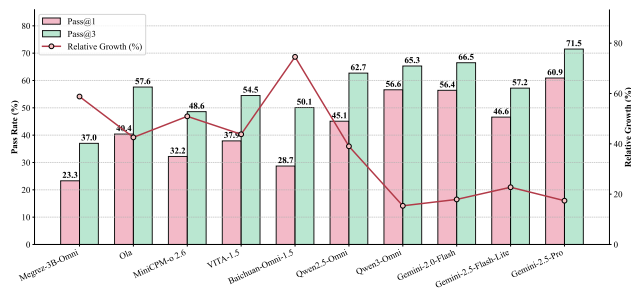


Figure 10. Comparison of Pass@1 and Pass@3 Scores.

By observing the results, we find that open-source Omni-MLLMs exhibit a larger performance gain when evaluated under the pass@3 metric, suggesting that they still have considerable potential to produce correct solutions through multiple attempts. In contrast, the closed-source Gemini models show relatively smaller improvements, indicating a higher degree of stability in their performance.

B.2.4. Results on Vision-Language Models

Besides the evaluation on Omni-MLLMs, we also conduct experiments on some Vision-Language Models. Since these models can not receive and understand audio input, we conduct three versions of experiments: (1) text and hand-drawn images from MathAll; (2) images from the original datasets with text from MathAll; and (3) text and images from the original datasets. Results are provided in Table 8.

Models	<u>T</u> & <u>I</u>	<u>T</u> & <u>I</u>	<u>T</u> & <u>I</u>
GPT-4V	29.4	32.0	33.7
GPT-4o	34.0	38.2	39.0
Qwen3-VL	57.2	62.1	68.5
Claude-Sonnet-4.5	56.6	62.0	68.9

Table 8. Results on Vision-Language Models across different input modalities. T (Text) and I (Image) represent results obtained from our MathAll benchmark, while the others correspond to results from the original datasets.

Vision-Language Models behave similarly to Omni-MLLMs, and their accuracy drops when evaluated with either text or image inputs in MathAll, indicating the challenge that MathAll presents to these models.

B.3. More Details and Analysis

Ablation Study Details. In our ablation studies, we evaluate six open-source and three closed-source Omni-MLLMs. However, due to the excessive time required by Qwen3-Omni to complete all experiments, we temporarily exclude it from this analysis. Nevertheless, we believe the results from the remaining models are sufficiently representative and provide strong evidence for our conclusions.

Analysis of CoT and Few-shot. In our experiments, we find that CoT and two-shot provide relatively little improvement in the models’ performance. Through a closer observation of the models’ generations, we find that this mainly occurs because of the lack of instruction-following abilities in open-source Omni-MLLMs. On the one hand, they may ignore instructions asking for a direct answer or step-by-step reasoning, and instead either generate only the final answer or produce detailed reasoning regardless of the prompts. On the other hand, they may struggle to understand what the few-shot examples are intended to teach, sometimes repeating the example answer while leaving the actual question unanswered. As for closed-source Omni-MLLMs, CoT can steadily improve their performance, while few-shot has little effect, which may be due to the diverse reasoning styles required for solving different questions.

C. Additional Analysis and Case Study

C.1. Representative Success Case

Some models demonstrate promising ability to handle information distributed across three modalities when solving some problems in MathAll. Figure 11 shows a representative success case from Qwen3-Omni. In this example, the model must jointly interpret the text, image, and audio inputs to construct a complete understanding of the problem. Qwen3-Omni uses the audio to confirm the shapes and positions of geometric objects in the image, exhibiting strong modality-complementary reasoning. After forming a coherent understanding, it correctly applies several geometric theorems and performs the necessary calculations step by step, reflecting its modality-integrated reasoning capability.

C.2. Overview of Error Types

In this subsection, we provide a detailed introduction to the different types of errors and illustrate each with a representative example.

Visual Perception Errors occur when the model fails to correctly extract information from the given image, such

as misreading textual details or misinterpreting angle correspondences. These errors are particularly likely to arise in our hand-drawn images. A typical example is shown in Figure 14, where Qwen3-Omni misidentifies the positions of the nail’s head and tail on the ruler, ultimately producing an incorrect answer.

Cross-Modal Integration Errors are a distinctive class of errors that become especially salient when evaluating models with our MathAll benchmark. They arise when the model fails to form a coherent understanding from information distributed across image+audio or text+image+audio modalities. This error type highlights insufficient modality-complementary reasoning ability. An illustrative case is provided in the main body of this paper.

Incorrect Solution Planning occurs when the model understands the given multimodal information but fails to devise an appropriate solution strategy. These errors reveal limitations in the model’s modality-integrated reasoning capabilities. As shown in Figure 15, Qwen3-Omni struggles to identify the pattern in the provided figures. It tries to solve it through trial-and-error, but ultimately fails to determine a correct solution path.

Contradictory Reasoning Errors also stem from weaknesses in modality-integrated reasoning. These errors arise when the model generates statements that contradict information it previously produced. This issue is likely related to the model’s limited attention span: mathematical reasoning often requires long, coherent chains of thought, yet during extended generation the model may lose track of earlier details. In the example presented in Figure 16, Qwen3-Omni initially identifies the grid as 3×4 , but later forgets this and incorrectly claims that a point lies at row 1, column 5.

Calculation Errors arise when the model makes mistakes in numerical computations. Since these models cannot rely on external tools such as calculators, all calculations must be performed through self-regressive prediction. Such errors commonly occur in floating-point and trigonometric operations. Figure 13 shows a typical case, where the model incorrectly computes 78.6×0.7660 , leading to an erroneous final answer.

C.3. Case Study of Gemini-2.5-Pro

Even though Gemini-2.5-Pro achieves the best performance on MathAll among all the Omni-MLLMs we evaluate, it still makes errors similar to Qwen3-Omni. In addition, it seems to face some unique challenges.

By looking closely at its generated answers, we find that compared with open-source models, it has difficulty extracting information accurately from the audio. This becomes a serious problem when the audio contains key conditions needed to solve the question. As shown in Figure 12, Gemini-2.5-Pro fails to receive the correct audio information and starts to hallucinate. In this example, it claims

that the audio says AD is parallel to BC and that angle C is 90 degrees, which are completely made up.

Because it relies on these incorrect details, Gemini-2.5-Pro cannot complete a proper reasoning path. After failing to continue, it starts to question the original problem conditions and introduces new assumptions on its own. These assumptions mostly come from the model’s past experience rather than the actual information provided. For instance, when it begins doubting the problem itself, it generates statements such as “Often in such problems, the goal is to find the tangent of one of the acute angles. Let’s assume the question is about the angle in a right triangle...”. Here the model uses its previous experience to reshape the problem instead of solving the one it is given. Furthermore, during the later reasoning steps, it continues to propose various assumptions about how the problem could be rewritten and tries to find a common answer based on those assumptions. These attempts do not help in solving the actual problem and only lead the reasoning further away from the correct solution.

To address these challenges, besides improving the model’s ability to understand audio, another possible approach is to encourage the model to acknowledge its uncertainty. In other words, when it finds that the problem is unsolvable for itself, it should be encouraged to say something like “I don’t know” instead of producing answers based on conditions that are not provided. This helps to avoid unnecessary resource consumption and prevents a further drop in the model’s faithfulness.

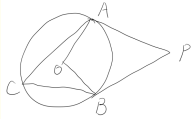
Success Case of Qwen3-Omni

Question: angle C = 65.0. then the degree of angle P is ()

Choices: A:65° B:130° C:50° D:100°

Audio: (🔊) Circle with center O has points A, B, and C on its circumference. Point P is outside the circle, and line segments PA and PB are tangent to the circle at points A and B, respectively.

Image:



Response from Qwen3-Omni:

Of course. Let's solve this geometry problem step-by-step using the information provided.

Given Information:

- We have a circle with center O.
- Points A, B, and C are on the circumference of the circle. Point P is outside the circle. Line segments PA and PB are tangent to the circle at points A and B, respectively.
- The measure of inscribed angle $\angle C$ is 65° .

Goal: Find the measure of $\angle P$

Step-by-Step Solution:

1. **Use the Inscribed Angle Theorem:** The inscribed angle theorem states that the measure of an inscribed angle is half the measure of its intercepted arc. The angle $\angle C$ intercepts the arc AB . Therefore: $\text{measure of arc } AB = 2 \times \angle C = 2 \times 65^\circ = 130^\circ$.

2. **Use the Tangent-Chord Theorem (Alternate Segment Theorem):** This theorem states that the angle between a tangent and a chord through the point of contact is equal to the inscribed angle in the alternate segment.

- The angle between the tangent PA and the chord AB is $\angle PAB$.

- This angle is equal to the inscribed angle subtended by arc AB in the alternate segment.

- The inscribed angle in the alternate segment is $\angle C$. Therefore: $\angle PAB = \angle C = 65^\circ$.

3. **Use the Tangent-Tangent Theorem:** Tangents drawn from an external point to a circle are equal in length. Therefore, $PA = PB$. Since $PA = PB$, the triangle $\triangle PAB$ is isosceles with base AB . In an isosceles triangle, the angles opposite the equal sides are equal. Therefore: $\angle PBA = \angle PAB = 65^\circ$.

4. **Find $\angle P$ using the Triangle Angle Sum Theorem:** The sum of the interior angles in any triangle is 180° .

In $\triangle PAB$: $\angle P + \angle PAB + \angle PBA = 180^\circ$

Substitute the known values: $\angle P + 65^\circ + 65^\circ = 180^\circ$, $\angle P + 130^\circ = 180^\circ$, $\angle P = 180^\circ - 130^\circ = 50^\circ$

Final Answer: The measure of angle P is 50° .

$\boxed{50^\circ}$

Answer: C

Figure 11. A typical success case of Qwen3-Omni from Plane Geometry on MathAll. Successful perceptions from the text, image and audio are highlighted in blue, yellow and green, respectively. The application of correct theorems is highlighted in purple.

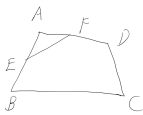
Failure Case of Gemini-2.5-Pro

Question: $EF = 2.0$. $BC = 5.0$. $CD = 3.0$. $\tan C$ is equal to ()

Choices: A: $\frac{3}{4}$ B: $\frac{4}{3}$ C: $\frac{3}{5}$ D: $\frac{4}{5}$

Audio: ♪) Quadrilateral ABCD has points E and F as midpoints on sides AB and AD, respectively, connected by segment EF.

Image:



Response from Gemini-2.5-Pro:

Here is the step-by-step solution:

1. **Analyze the provided information:** We are given a quadrilateral ABCD. The audio description specifies that this is a trapezoid **where AD is parallel to BC** ($AD \parallel BC$). The audio also states that **$\angle C = 90$ degrees**. This makes the trapezoid a right trapezoid. We are given the lengths of several segments: $EF = 2.0$, $BC = 5.0$, and $CD = 3.0$. The points E and F are the midpoints of the non-parallel sides AB and CD, respectively. The segment EF is the median (or midsegment) of the trapezoid.

...

7. **Analyze the result:** A side length cannot be negative. This indicates a contradiction in the given values. The problem as stated is impossible. Let's review the information again.

8. **Re-evaluating the Problem:** **Let's check the possibility that the parallel sides are AB and DC, or that the right angle is at B instead of C.** Let's reconsider the initial interpretation. Perhaps the drawing is misleading, and the question is simpler. Let's assume there is a typo in the provided numbers and try to work backwards from the choices.

9. **A common type of problem:** **Often in such problems, the goal is to find the tangent of one of the acute angles. Let's assume the question is about an angle in a right triangle whose sides are given by the numbers in the problem.**

... **Let's assume the right angle is at D.** Then we have a right trapezoid with $AD \parallel BC$ and $\angle D = 90^\circ$. We would also need $\angle A = 90^\circ$. This would make it a rectangle, which doesn't fit the drawing.

... **Let's assume the right angle is at B.**

... **Let's assume the problem intended to define a right triangle with sides that could be found from the trapezoid.** If we draw a perpendicular from D to BC at point H. If we ignore the given $EF=2$ and assume $BH=4$ (to make a 3-4-5 triangle), then $\tan(B) = DH/BH = 3/4$. This matches option A.

... **Let's assume there is a right triangle where $CD=3$ is one leg and some other segment is the other leg.** Let's assume the question implicitly defines a right triangle where the adjacent side to angle C is 4 and the opposite side is 3.

...

Answer: B

Figure 12. A typical failure case of Gemini-2.5-Pro from Plane Geometry on MathAll, with the error highlighted in red and incorrect assumptions are highlighted in orange.

Calculation Errors

Question: Calculate the area of the following triangle. Round your answer to two decimal places.

Audio: (🔊) A triangle has one angle of 50 degrees, with sides measuring 12 meters and 13.1 meters.

Image:



Response from Qwen3-Omni:

Based on the information provided, we have a triangle with two sides and the included angle. We can use the formula for the area of a triangle given two sides and the included angle.

The formula is: $A = \frac{1}{2}ab \sin(C)$ Where a and b are the lengths of two sides and C is the measure of the included angle.

From the diagram and the problem statement: $a = 12$ m, $b = 13.1$ m, $C = 50^\circ$. Now, substitute these values into the formula:

$$A = \frac{1}{2} \times 12 \times 13.1 \times \sin(50^\circ)$$

First, calculate the product of the sides: $12 \times 13.1 = 157.2$

Now, multiply by $\frac{1}{2}$: $\frac{1}{2} \times 157.2 = 78.6$

Next, find the sine of 50° : $\sin(50^\circ) \approx 0.7660$

Now, multiply the two results: $A \approx 78.6 \times 0.7660 \approx 60.1176$

Finally, round the answer to two decimal places: $A \approx 60.12 \text{ m}^2$

The area of the triangle is approximately **60.12** square meters.

Answer: 60.21

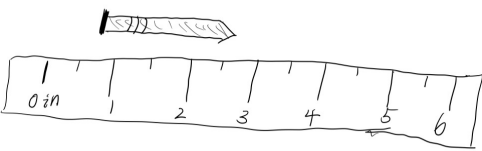
Figure 13. A typical calculation error of Qwen3-Omni from **Plane Geometry** on MathAll, with the error highlighted in red.

Visual Perception Errors

Question: The nail is about () inches long.

Audio: (🔊) The image displays a nail positioned above a ruler, which is marked from 0 to 6.

Image:



Response from Qwen3-Omni:

Based on the provided image, we can determine the length of the nail by reading the ruler.

1. Identify the starting point: **The left end of the nail is aligned with the '0 in' mark on the ruler.**

2. Identify the ending point: **The right end of the nail extends past the '4' mark and stops at the mark that is one small tick after '4'.**

3. Read the measurement: The ruler is marked in inches. The small tick marks between the whole numbers represent fractions of an inch. **Since there are 4 small ticks between each inch, each tick represents $\frac{1}{4}$ of an inch.** The mark at '4' is 4 inches. The first tick after '4' is $4 + \frac{1}{4} = 4\frac{1}{4}$ inches. The second tick is $4 + \frac{2}{4} = 4\frac{2}{4}$ inches (or $4\frac{1}{2}$ inches). The third tick is $4 + \frac{3}{4} = 4\frac{3}{4}$ inches. The fourth tick is $4 + \frac{4}{4} = 5$ inches.

The right end of the nail aligns with the second small tick mark after the '4' mark. Therefore, the length of the nail is $4\frac{2}{4}$ inches, which simplifies to $4\frac{1}{2}$ inches. The nail is about $4\frac{1}{2}$ inches long.

Answer: $4\frac{1}{2}$

Answer: 2

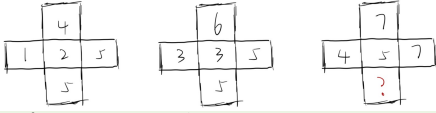
Figure 14. A typical visual perception error of Qwen3-Omni from **Others** on MathAll, with the error highlighted in red.

Incorrect Solution Planning

Question: Which number is missing?

Audio: (1) The first figure contains the numbers 4, 1, 2, 5, and 5, with 2 in the center. The second figure has the numbers 6, 3, 3, 5, and 5, with 3 in the center. The third figure includes the numbers 7, 4, 5, 7, and a question mark, with 5 in the center. The question mark is at the bottom.

Image:



Response from Qwen3-Omni:

Let's analyze the pattern in the figures. Each figure is a cross shape with five numbers: One number in the center. One number above the center. One number below the center. One number to the left of the center. One number to the right of the center. Let's denote the numbers as: Top: \$T\$, Left: \$L\$, Center: \$C\$, Right: \$R\$, Bottom: \$B\$. We need to find a relationship between these numbers.

Let's look at the relationship between the numbers. Let's try to see if there's a pattern connecting the center number \$C\$ to the others. Looking at the first figure: \$T = 4\$, \$L = 1\$, \$C = 2\$, \$R = 5\$, \$B = 5\$. Notice that the numbers on the left and right are 1 and 5, and the center is 2. The top is 4, and the bottom is 5. Looking at the second figure: \$T = 6\$, \$L = 3\$, \$C = 3\$, \$R = 5\$, \$B = 5\$. The center is 3, and the left number is also 3. Looking at the third figure: \$T = 7\$, \$L = 4\$, \$C = 5\$, \$R = 7\$, \$B = ?\$. The center is 5, and the right number is also 7. The top is 7. Wait, let's look at the differences.

Looking at the first figure: \$T - L = 4 - 1 = 3\$, \$R - C = 5 - 2 = 3\$, \$B - C = 5 - 2 = 3\$. Hmm, interesting. \$T - L = R - C = B - C = 3\$. Looking at the second figure: \$T - L = 6 - 3 = 3\$, \$R - C = 5 - 3 = 2\$, \$B - C = 5 - 3 = 2\$. Not the same. Looking at the first figure again: \$T - C = 4 - 2 = 2\$, \$C - L = 2 - 1 = 1\$, \$C - B = 2 - 5 = -3\$. No clear pattern. Wait, let's look at the sum. First figure: Sum = \$4 + 1 + 2 + 5 + 5 = 17\$. Second figure: Sum = \$6 + 3 + 3 + 5 + 5 = 22\$. Third figure: Sum = \$7 + 4 + 5 + 7 + ? = 23 + ?\$. Not obvious. Another idea: perhaps the center number is the average of the other numbers? First figure: Average of \$T, L, R, B = (4 + 1 + 5 + 5)/4 = 15/4 = 3.75\$, not 2. No. Second figure: Average of \$T, L, R, B = (6 + 3 + 5 + 5)/4 = 19/4 = 4.75\$, not 3. No. Another idea: perhaps the center number is related to the left and right numbers. First figure: \$L = 1\$, \$R = 5\$, \$C = 2\$. \$C = (L + R)/3 = (1 + 5)/3 = 2\$. Oh! That's 2. Second figure: \$L = 3\$, \$R = 5\$, \$C = 3\$. \$(L + R)/3 = (3 + 5)/3 = 8/3 \approx 2.67\$, not 3. No. Wait, maybe \$C = (L + R)/2\$? First: \$(1 + 5)/2 = 3\$, not 2. Second: \$(3 + 5)/2 = 4\$, not 3.

No. Another idea: perhaps the sum of the...

Answer: 2

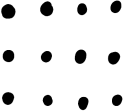
Figure 15. A typical incorrect solution planning error of Qwen3-Omni from Others on MathAll, with the error highlighted in red.

Contradictory Reasoning Errors

Question: We take three points from the grid so that they are collinear. How many possibilities do we have?

Audio: (1) A grid of 12 points arranged in three rows and four columns.

Image:



Response from Qwen3-Omni:

Of course, let's solve this step by step. The problem asks for the number of ways to choose three collinear points from a 3x4 grid. The grid has 3 rows and 4 columns, for a total of 12 points. To find the number of sets of three collinear points, we need to consider all possible straight lines that pass through at least three points in the grid. We can categorize these lines based on their orientation: 1. Horizontal Lines 2. Vertical Lines 3. Diagonal Lines (with positive and negative slopes)

1. Horizontal Lines Each row contains 4 points. We need to find how many ways we can choose 3 collinear points from a row of 4. This is a combination problem. The number of ways to choose 3 items from 4 is given by the combination formula: \$C(n, k) = \frac{n!}{k!(n-k)!}\$. For a row of 4 points, the number of ways to choose 3 is: \$C(4, 3) = \frac{4!}{3!(4-3)!} = \frac{4 \times 3!}{3! \times 1!} = 4\$. Since there are 3 rows, the total number of horizontal sets of 3 collinear points is: \$3 \times 4 = 12\$.

2. Vertical Lines

Each column contains 3 points. We need to find how many ways we can choose 3 collinear points from a column of 3. The number of ways to choose 3 items from 3 is: \$C(3, 3) = \frac{3!}{3!(3-3)!} = \frac{3!}{3! \times 0!} = 1\$. Since there are 4 columns, the total number of vertical sets of 3 collinear points is: \$4 \times 1 = 4\$.

3. Diagonal Lines

Now we need to find sets of 3 collinear points that are not horizontal or vertical. These will be on diagonal lines with a slope other than 0 or infinity. Let's consider the possible slopes. Diagonals with slope = 1 (rising from left to right) We can try to find diagonals that contain at least 3 points. Starting from the top-left corner (row 1, col 1): The points would be (1,1), (2,2), (3,3). This is a diagonal of length 3. Number of ways to choose 3 points from these 3: \$C(3,3) = 1\$. Starting from (1,2): The points would be (1,2), (2,3), (3,4). This is another diagonal of length 3. Number of ways: \$C(3,3) = 1\$. Starting from (1,3): The next point would be (2,4). There is no point at (3,5). So this diagonal only has 2 points. Starting from (2,1): Only one point. Starting from (2,2): The next point would be (3,3), and the previous point would be (1,1). This is the same diagonal as the first one we found. So, there are only two diagonals with slope 1 that contain exactly 3 points. This gives us 2 sets.

Diagonals with slope = -1 (falling from left to right) Starting from the top-right corner (1,4): The points would be (1,4), (2,3), (3,2). This is a diagonal of length 3. Number of ways: \$C(3,3) = 1\$. Starting from (1,3): The points would be (1,3), (2,2), (3,1). This is another diagonal of length 3. Number of ways: \$C(3,3) = 1\$. Starting from (1,2): The next point would be (2,1). There is no point at (3,0). So only 2 points. Starting from (1,1): Only one point. Starting from (2,4): The next point would be (3,3), and the previous point would be (1,5). So only 2 points.

Answer: 20

Figure 16. A typical contradictory reasoning error of Qwen3-Omni from Others on MathAll, with the error highlighted in red.

Model	Generation Type	Generation Settings
Qwen3-Omni	Direct & 2-shot CoT 0-shot CoT	temperature = 0.7, max_tokens = 512 temperature = 0.7, max_tokens = 512
Qwen2.5-Omni	Direct & 2-shot CoT 0-shot CoT	temperature = 0.7, max_tokens = 256 temperature = 0.7, max_tokens = 256
Baichuan-Omni-1.5	Direct & 2-shot CoT 0-shot CoT	temperature = 0.7, max_tokens = 1024 temperature = 0.7, max_tokens = 1024
MiniCPM-o 2.6	Direct & 2-shot CoT 0-shot CoT	temperature = 0.7, max_tokens = 512 temperature = 0.7, max_tokens = 512
VITA-1.5	Direct & 2-shot CoT 0-shot CoT	temperature = 0.7, max_tokens = 512 temperature = 0.7, max_tokens = 512
Ola	Direct & 2-shot CoT 0-shot CoT	temperature = 0.7, max_tokens = 1024 temperature = 0.7, max_tokens = 1024
Megrez-3B-Omni	Direct & 2-shot CoT 0-shot CoT	temperature = 0, max_tokens = 512 temperature = 0.7, max_tokens = 512
Gemini-2.0-Flash	Direct & 2-shot CoT 0-shot CoT	temperature = 0, max_tokens = 2048 temperature = 0.7, max_tokens = 2048
Gemini-2.5-Flash-Lite	Direct & 2-shot CoT 0-shot CoT	temperature = 0, max_tokens = 2048 temperature = 0.7, max_tokens = 2048
Gemini-2.5-Pro	Direct & 2-shot CoT 0-shot CoT	temperature = 0, max_tokens = 2048 temperature = 0.7, max_tokens = 2048

Table 9. Generation hyperparameters used for all models evaluated in our benchmark.

Problem Text Split Prompt

You will be given a geometry problem with two inputs:

1. An image (the figure with labels).
2. A problem text (the question and additional conditions).

Your task is to split the information in the problem text into three categories:

- Descriptive Information: Information that can be directly observed in the image, such as labeled angle measures (e.g., $\angle ABC = 30^\circ$), segment lengths shown on the figure (e.g., $AB = 5$), parallel/perpendicular/intersecting relationships, and points or lines explicitly marked in the figure.
- Supplementary Information: Information provided in the problem text but not explicitly marked in the image. Examples include exact values of angles not shown in the figure, conditions like "AB = 6 cm" without being labeled, algebraic relations, or extra constraints introduced in the text.
- Question: The actual question being asked, such as "then angle ACD = () Choices: A:120° B:130° C:140° D:150°".

Rules:

- If the information is visible in the image, classify it as Descriptive Information.
- If the information appears only in the text and not explicitly in the image, classify it as Supplementary Information.
- The Question should be part of the original problem text.
- **Preserve the original mathematical notation style (symbols, LaTeX, degree marks, etc.) in the Supplementary Information and Question sections.**
- **In the Descriptive Information section, convert mathematical notations (symbols, LaTeX, degree marks) into natural language equivalents. For example: "30°" → "30 degrees", " $\angle ABC$ " → "angle ABC", " $x = 5$ " → "x equals 5".**
- If the original problem text is in Chinese, output all sections (Descriptive Information, Supplementary Information, and Question) in Chinese.

The given problem text is: {text}.

Output strictly in JSON format:

```
{
  "Descriptive Information": ["...", "..."],
  "Supplementary Information": ["...", "..."],
  "Question": "..."
}
```

Question Consistency Check Prompt

Task: You will be given two statements (Statement A and Statement B). Each statement asks a question. Your job is to determine whether both statements are asking the same question. They may use different wording or include extra details, but if they ultimately ask for the same information or have the same intent, treat them as equivalent.

Output rules:

1. If they ask the same question (same core intent or target answer), output 1.
2. If they ask different questions (different goals, answer types, or focuses), output 0.
3. Output only "1" or "0" with no explanations.

Table 10. Prompts used in creating **Problem Text**.

Captioning Prompt

You are given a mathematical figure. Your task is to generate a concise and accurate description of the figure.

The priority of your description is:

1. Accuracy: Ensure every detail you describe can be directly verified from the figure.
2. Completeness: After accuracy, try to include the most important elements, but do not sacrifice accuracy for completeness.

You will also be provided with some known reference information.

- Do not repeat this reference information in your output.
- You may use it as context to better understand the figure, but the description should focus only on what can be visually confirmed from the figure.

Your description should highlight the key elements of the figure, such as:

- Points, lines, and segments explicitly labeled.
- Marked angles or angle measures.
- Marked lengths, parallel or perpendicular relationships.
- Intersections or special configurations clearly visible.

Avoid unnecessary speculation or rephrasing of the reference information.

Additional Rules:

- If the Reference Information is in Chinese, output the entire description in Chinese, else in English.
- For chart or graph-type figures, explicitly extract and include coordinate axis labels, scales, and title information if they are clearly visible.

Input:

- Figure
- Reference Information: {reference_text}

Output:

A short, precise description of the key information visible in the figure.

Common Information Extraction Prompt

You are given two text descriptions of a mathematical figure. Your task is to identify and extract the common elements that are present in both descriptions.

Please provide a concise summary that includes only the information that is explicitly mentioned in both texts. Avoid adding any new information or interpretations.

****Additional Requirement:****

- Output only the common elements themselves. Do not include explanations, formatting text, or any other output beyond the common elements.
- Present the common elements in clear, natural language so that the output is directly readable and understandable.

Text 1: {text1}

Text 2: {text2}

Output: (Common elements only.)

Table 11. Prompts used in creating **Explanatory Audio**.

Deduplication and Fluency Optimization Prompt

Task: You are given a piece of text that may contain redundant or repetitive information. Your job is to perform two steps:

1. Remove all repeated or redundant information while keeping the remaining content factually correct and coherent.
2. Rewrite the remaining information in clear, natural, and fluent English so that it can be read aloud smoothly as a natural paragraph.

Requirements:

1. Accuracy is the top priority.
 2. If you are not sure about the meaning of certain parts of the text, you may omit them rather than make assumptions.
 3. The final output should be concise, logically organized, and easy to understand when spoken aloud.
 4. Directly output only the processed text without any explanations or additional comments.
-

Audio Text Consistency Check Prompt

Task: You will be given two statements (Statement A and Statement B). Your job is to determine whether both statements contain the same information.

They may use different wording or sentence structure, but if they express the same meaning or convey the same facts, they should be considered equivalent.

Output rules:

1. If both statements express the same meaning or provide the same information using the same language, output 1.
2. If they express different meanings or contain different information or use different languages, output 0.

Output only “1” or “0” with no explanations.

Table 12. Prompts used in refining **Explanatory Audio**.

Modal	Prompt
Hand-drawn image+Text+Audio for Direct	<p>You are a multimodal reasoning model. You will receive a math problem in text form, along with a hand-drawn image and an audio description that supplements the diagram.</p> <p>Please solve the problem based on the provided input and directly output the final answer in one "□". Do not explain your reasoning or show intermediate steps. If it is a multiple choice question, only one letter is allowed in the the "□". The math problem is:</p>
Hand-drawn image+Text+Audio for CoT	<p>You are a multimodal reasoning model. You will receive a math problem in text form, along with a hand-drawn image and an audio description that supplements the diagram.</p> <p>Please solve the problem step by step and put your answer in one "□". If it is a multiple choice question, only one letter is allowed in the "□". The math problem is:</p>
Hand-drawn image+Text for CoT	<p>You are a multimodal reasoning model. You will receive a math problem in text form, along with a hand-drawn image.</p> <p>Please solve the problem step by step and put your answer in one "□". If it is a multiple choice question, only one letter is allowed in the "□". The math problem is:</p>
Text+Audio for CoT	<p>You are a multimodal reasoning model. You will receive a math problem in text form, along with an audio description.</p> <p>Please solve the problem step by step and put your answer in one "□". If it is a multiple choice question, only one letter is allowed in the "□". The math problem is:</p>
Raw image+Text for CoT	<p>You are a multimodal reasoning model. You will receive a math problem in text form, along with an image.</p> <p>Please solve the problem step by step and put your answer in one "□". If it is a multiple choice question, only one letter is allowed in the "□". The math problem is:</p>

Table 13. The prompts used in the evaluation for response generation.

CoT with 2-shot Examples

You are a multimodal reasoning assistant specialized in solving math problems. You will receive a math problem in text form, along with a hand-drawn image and an audio description that supplements the diagram.

Please solve the problem step by step and put your answer in one "". If it is a multiple choice question, only one letter is allowed in the "".

Example 1

[Text]:

Find x .

[Image]:

[A hand-drawn image that looks like a right triangle]

[Audio]:

[A man is saying: "In this right triangle, one leg measures 20 units and the other leg measures 22 units."]

Output:

We have a right triangle with the two legs: First leg = 20, second leg = 22. We want to find the hypotenuse.

Step 1: Use the Pythagorean theorem

For a right triangle, the square of the hypotenuse equals the sum of the squares of the two legs:

$$c^2 = a^2 + b^2$$

Step 2: Substitute the values

$$c^2 = 20^2 + 22^2$$

Step 3: Calculate the squares

$$20^2 = 400, 22^2 = 484$$

Step 4: Add the squares

$$c^2 = 400 + 484 = 884$$

Step 5: Take the square root

$$c = \sqrt{884}$$

c can be simplified as $2\sqrt{221}$. So the answer is $2\sqrt{221}$.

Example 2

[Text]:

If line WP is a perpendicular bisector, $\angle WHA = 8q + 17$, $\angle HWP = 10 + q$, $AP = 6r + 4$, and $PH = 22 + 3r$, find q . Choices: A:5 B:6 C:7 D:8

[Image]:

[A hand-drawn image that shows a triangle with points W , H , A , and P labeled accordingly. Point X is inside the triangle.]

[Audio]:

[A woman is saying: "The image shows a triangle labeled WHA , with two points P and Q on sides HA and WA , respectively. A line segment WX is drawn inside the triangle, connecting vertex W to a point X on line segment HP ."]

Output:

Since WP is a perpendicular bisector of HA , $\angle HWP + \angle WHA = 90^\circ$. Given $\angle HWP = 10 + q$, $\angle WHA = 8q + 17 \Rightarrow (10 + q) + (8q + 17) = 90 \Rightarrow 9q + 27 = 90 \Rightarrow 9q = 63 \Rightarrow q = 7$.

So the answer is C .

Now solve the following problem:

[Text]: {question_text}

[Image]: Provided.

[Audio]: Provided.

Table 14. Two-shot CoT prompt used in the evaluation for response generation.

Answer Extraction Prompt

I am providing you a response from a model to a math problem, termed 'Model Response'. You should extract the answer from the response as 'Extracted Answer'. Directly output the extracted answer with no explanation.

1. Model response: 'Rounded to two decimal places, the perimeter of the sector is approximately: (-2, 1)'

Extracted Answer: (-2, 1)

2. Model response: 'at those points. Therefore, the correct option that represents the meaning of the intersection points of the graphs is: D. They give the solutions to the equation $f(t) = g(t)$.'

Extracted Answer: D

3. Model response: 'at 1 (there's a closed circle at $y = 1$), the range in interval notation is $((-4, 1])$. Final values: Domain: $((-3, 3])$, Range: $((-4, 1])$ '

Extracted Answer: Domain: $((-3, 3])$, Range: $((-4, 1])$

4. Model response: 'As it stands, I cannot provide the correct option letter because there isn't enough information to solve for 'y'.'

Extracted Answer: null

5. Model response: 'Given that $AB = 17.6$ meters, we can now substitute into the equation: $d = 17.6 / \cos(38^\circ)$. Therefore, to one decimal place, the distance d between Ned and Bart is approximately 22.3 meters.'

Extracted answer: 22.3

6. Model response: 'The model gives the coefficients for a quadratic function, $f(x) = ax^2 + bx + c$, and specifically the graphed function is $f(x) = -x^2 - 2x + 1$.'

Extracted answer: $f(x) = -x^2 - 2x + 1$.

Table 15. Answer extraction prompt used in the evaluation.

Answer Comparison Prompt

Below are two answers to a math question. Question is [Question], [Standard Answer] is the standard answer to the question, and [Model_answer] is the answer extracted from a model's output to this question. Determine whether these two answers are consistent.

Please note that only when the [Model_answer] completely matches the [Standard Answer] means they are consistent. For non-multiple-choice questions, if the meaning is expressed in the same way, it is also considered consistent, for example, 0.5m and 50cm. For multiple-choice questions, if the [Model_answer] matches either the option letter or the full text of the correct choice, they are considered consistent. If they are consistent, Judgement is 1; if they are different or the model answer is unfinished, Judgement is 0.

[Question]: Write the set of numbers represented on the number line in interval notation.

[Standard Answer]: (-2,1]

[Model_answer] : Extracted Answer: ((-2, 1))

Judgement: 0

[Question]: As shown in the figure, circle O has a radius 1.0, if angle BAC = 60.0, then the length of BC is () Choices:A:2 B:2\sqrt{3} C: \sqrt{3} D:2\sqrt{2}

[Standard Answer]: C

[Model_answer] : B:2 \sqrt{3} Judgement: 0

[Question]: Find the domain and range of the function f using interval notation.

[Standard Answer]: domain: [-4, 0) and range: (-3, 1]

[Model_answer] : Range: ((-4, 1))

Judgement: 0

[Question]: As shown in the figure, circle O has a radius 1.0, if angle BAC = 60.0, then the length of BC is () Choices: A:2 B:2\sqrt{3} C:\sqrt{3} D:2\sqrt{2}

[Standard Answer]: C

[Model_answer] : null

Judgement: 0

[Question]: Given the graph of the ellipse that intersects with x-axis at 9 and -9 and with y-axis at 3 and -3, determine its equation.A. $\frac{x^2}{81} + \frac{y^2}{9} = 1$. B. Can not determine.

[Standard Answer]: A

[Model_answer] : $\frac{x^2}{81} + \frac{y^2}{9} = 1$

Judgement: 1

[Question]: What is $\angle BAC$, in degrees?

[Standard Answer]: 55

[Model_answer] : Can you provide more information? If you can, I will solve this problem.

Judgement: 0

[Question]: {question}

[Standard Answer]: {gt}

[Model_answer] : {extraction}

Judgement:

Table 16. Answer comparison prompt used in the evaluation.