MathSight: A Benchmark Exploring Have Vision-Language Models Really Seen in University-Level Mathematical Reasoning?

Anonymous ACL submission

Abstract

Recent advances in Vision-Language Models (VLMs) have achieved impressive progress in multimodal mathematical reasoning. Yet, how much visual information truly contributes to reasoning remains unclear. Existing benchmarks report strong overall performance but seldom isolate the role of the image modality, leaving open whether VLMs genuinely leverage visual understanding or merely depend on linguistic priors. To address this, we present **MathSight**, a university-level multimodal mathematical reasoning benchmark designed to disentangle and quantify the effect of visual input. Each problem includes multiple visual variants—original, hand-drawn, photo-captured—and a text-only condition for controlled comparison. Experiments on state-of-the-art VLMs reveal a consistent trend: the contribution of visual information diminishes with increasing problem difficulty. Remarkably, Qwen3-VL without any image input surpasses both its multimodal variants and GPT-5, underscoring the need for benchmarks like MathSight to advance genuine vision-grounded reasoning in future models.

1 Introduction

Recent advances in Vision-Language Models (VLMs) have led to significant progress in multimodal reasoning, particularly in mathematical problem solving. A growing collection of benchmarks, such as MathVista (Lu et al., 2023), MathVerse (Zhang et al., 2024b), MATH-Vision (Wang et al., 2024), MathCheck (Zhou et al., 2024), and Dynamath (Zou et al., 2024), has catalyzed rapid progress in this field, driving models toward increasingly sophisticated multimodal reasoning capabilities.

Despite these advancements, the role of visual information in mathematical reasoning remains both underexplored and ambiguous. While existing benchmarks report impressive overall scores, they rarely isolate how much the image modality actually contributes to reasoning. As a result, it is still unclear whether current VLMs genuinely perform vision-grounded reasoning or merely rely on textual priors to approximate correct answers. As shown in Figure 1, existing multimodal mathematical reasoning benchmarks typically include only a single, clean visual representation for each problem. However, after intentionally introducing additional visual variants, including *photo-captured* and *hand-drawn* versions, we find that many problems that models previously answered correctly became incorrect. This observation reveals that even subtle changes in visual appearance can disrupt the reasoning process of current VLMs, suggesting that their "visual understanding" may be superficial and overly dependent on presentation style rather than underlying geometry or semantics. This motivates our systematic investigation of how visual variation affects reasoning stability and accuracy.

To address this issue, we introduce **MathSight**, a new university-level multimodal mathematical reasoning benchmark designed to disentangle and quantify the influence of visual information. Each problem instance is presented under multiple visual variants, such as the original figure, a hand-drawn version, and a photo-captured version, enabling a controlled and systematic analysis of how variations in visual representation impact reasoning. In addition, we incorporate a text-only condition without any image input to further evaluate whether VLMs can solve problems purely through linguistic reasoning, independent of visual cues. Moreover, we also construct a university-level text-only benchmark to assess the reasoning capability of large language models (LLMs) under comparable difficulty settings, serving as a complementary reference for the multimodal analysis.

Through extensive experiments on multiple state-of-the-art VLMs (e.g., Gemini-2.5-pro, GPT-5, Claude-Opus, Claude-Sonnet, and Qwen3-vl), we

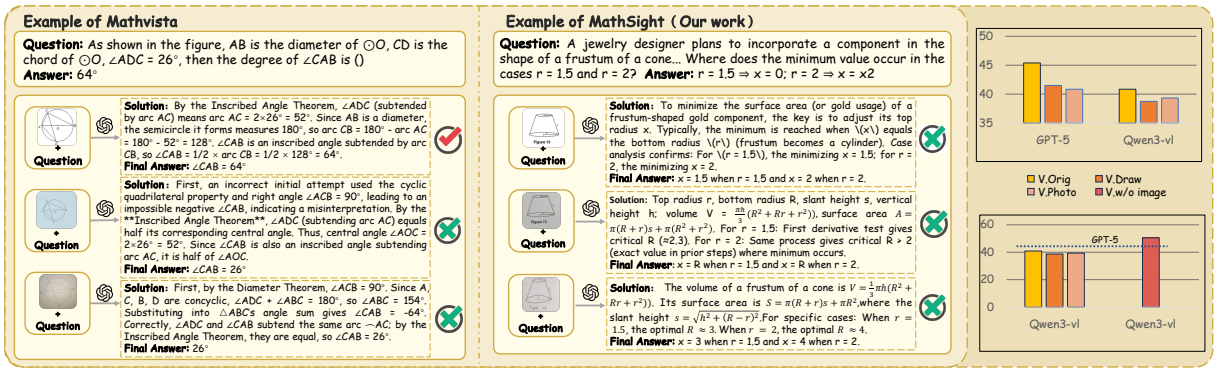


Figure 1: Motivation behind **MathSight**.

uncover a revealing trend: the contribution of visual information to mathematical reasoning diminishes as problem difficulty increases. In our university-level MathSight benchmark, the three visual variants—original, hand-drawn, and photo-captured variants—yield no statistically significant differences in model performance (See Figure 1). Even more intriguingly, Qwen3-VL achieves 50.53% accuracy without any image input, surpassing not only its own multimodal variants (*original*: 40.85, *hand-drawn*: 38.73, *photo-captured*: 39.33) but also outperforming GPT-5 (45.39) under identical conditions. This striking result indicates that high-level reasoning in current VLMs may rely more on textual priors and linguistic abstraction than on genuinely grounded visual understanding.

The main contributions of this paper are summarized as follows:

- **A new benchmark for vision-grounded mathematical reasoning.** We introduce **MathSight**, a university-level multimodal mathematical reasoning benchmark specifically designed to disentangle and quantify the influence of visual information. It enables systematic evaluation of VLMs under multiple controlled visual conditions.

- **A comprehensive analysis of visual sensitivity in VLMs.** By evaluating several state-of-the-art models (e.g., Gemini 2.5 Pro, GPT-5, Claude-Opus, Claude-Sonnet, and Qwen3-VL) across original, hand-drawn, photo-captured, and text-only settings, we reveal that the impact of visual input decreases as the problem difficulty increasing.

- **Empirical insights into the limits of current multimodal reasoning.** Our findings show that models often achieve comparable or even superior performance without image input—highlighting a fundamental gap between true visual understanding and language-driven reasoning in current VLMs.

2 MathSight

In this section, we introduce **MathSight**, a benchmark designed to investigate how visual information influences mathematical reasoning in vision-language models (VLMs). By varying the visual input—through original, hand-drawn, and photo-captured figures—while keeping the problems fixed, MathSight enables controlled analysis of visual contributions to reasoning performance and provides new insights into multimodal understanding. In the following, we describe the dataset construction, dataset composition, data analysis, and metrics for the logical consistency of MathSight in detail.

2.1 Dataset Construction

The pipeline of dataset construction includes three main phases: document selection, QA pair generation, and image augmentation.

2.1.1 Document Selection.

The dataset was collected from proprietary, non-public PDF documents provided by a collaborating company. These documents have never been made publicly available, ensuring that no evaluated model has been exposed to the data during pretraining.

We performed multiple rounds of manual screening. First, non-mathematical documents and those below the university level were excluded. Then, domain knowledge was utilized to classify and curate high-quality problem sets. Only problems contain-

Table 1: Existing Benchmarks, where "Visual.Variant" indicates that visual input has different variants, "Grad. Level" presents graduate-level problems and "Pro.Q" stands for proving questions, i.e., questions that cannot be simply verified against a standard answer.

Benchmark	Grad. Level	Pro.Q	Visual.Variant			Image.Size
			Original	Hand-drawn	Photo-captured	
TheoremQA (Chen et al., 2023)	✗	✗	✓	✗	✗	✗
MathVista (Lu et al., 2023)	✗	✗	✓	✗	✗	✗
Scibench (Wang et al., 2023)	✗	✗	✓	✗	✗	✗
QRData (Liu et al., 2024b)	✗	✗	✓	✗	✗	✗
MATH-Vision (Wang et al., 2024)	✗	✗	✓	✗	✗	✗
$MMMU_{math}$(Yue et al., 2024)	✗	✗	✓	✗	✗	✗
U-Math (Chernyshev et al., 2024)	✗	✗	✓	✗	✗	✗
Dynamath (Zou et al., 2024)	✗	✗	✓	✗	✗	✗
MathCheck (Zhou et al., 2024)	✗	✗	✓	✗	✗	✗
MathSight (our work)	✓	✓	✓	✓	✓	✓

ing a complete `question`, `answer/solution`, and `groundtruth` were retained.

2.1.2 QA Pair Generation.

After filtering the PDFs, we used `GPT-4o` to extract problems and convert them into our standardized `JSONL` format. A reusable extraction strategy was designed to guide the model, which can consistently produce well-formatted outputs.

The strategy consists of two parts: 1) a standard formatting guide specifying required fields (`question`, `answer/solution`, `groundtruth`, and optional `image_path`), and 2) error-handling rules such as skipping incomplete items and logging issues. In most cases, each multimodal problem is associated with a unique image. A few exceptions exist where multiple questions share a single visual reference.

Following extraction, we used `GPT-4o` again to flag questions below the university level. These were re-checked by human experts. Only problems verified to meet undergraduate-level or graduate-level difficulty were retained.

2.1.3 Image Augmentation.

Last but not least, we conduct augmentation for the images in multimodal problems manually. In order to simulate a real situation, we draw all the images with 5 graduate students whose handwriting is different from one another. Then the hand-drawn figures are taken as photos to use. To imitate photos that people take from paper materials, we print all the images and collect the photo-captured version of the image. Then the two variants of the original image possess large sizes for use with a mobile phone camera. These raw photos of variants are cropped and reduced in size to keep consistent with the original images.

2.2 Data Analysis

This section exposes the data analysis of MathSight from different perspectives.

2.2.1 Multimodal Coverage.

Our dataset contains 661 university-level multimodal problems, among which 603 are at the graduate-level, This number at the graduate level significantly exceeds that of other benchmarks. The right part of Figure 1 presents a representative multimodal example in MathSight, which is a geometric problem-solving task involving hypothesis reasoning and spatial visualization. For such problems, a model is considered correct only if it provides the exact final answer, accompanied by a logically consistent reasoning process that correctly understands the given graph. Partial answers, incorrect intermediate steps, or misinterpretations of the visual information may lead to wrong answers.

2.2.2 Subject Distribution.

We construct university-level problems across six core mathematical domains: Calculus, Algebra, Analysis, Probability & Statistics, Discrete Mathematics, and Applied Math. These domains encompass both foundational undergraduate courses and more advanced or specialized areas, such as mathematical logic and topology, which are essential for assessing deep mathematical reasoning capabilities.

Calculus and Applied Math constitute the largest portions of the dataset, accounting for 82.15% and 11.8%. The remaining domains contribute

meaningfully to the coverage. The distribution of undergraduate-level and graduate-level problems within each subject allows us to analyze the models' capabilities across varying levels. These distributions would help provide valuable insight into the areas where current large models excel or struggle.

2.2.3 Proving Problems.

Our dataset contains a total of 30 proving problems, accounting for 4.54% of all multimodal questions. Among them, 29 proving problems are at the graduate level, while only one is at the undergraduate level. We intentionally searched proving problems for MathSight because existing benchmarks exhibit deficiencies on such tasks. The performances of large models on these tasks are worth exploring. Moreover, proving problems are distributed across all six domains, enabling us to evaluate models' reasoning abilities from multiple disciplinary perspectives.

2.3 Metrics for Logical Consistency

MathSight dataset includes a considerable proportion of proving problems. To evaluate the logical reasoning of large models on these problems, we propose several metrics inspired by recent studies (Fu et al., 2025; Kang et al., 2025).

Token Confidence. Following (Fu et al., 2025), the token confidence C_i at position i is defined as

$$C_i = -\frac{1}{k}\sum_{j=1}^{k}\log P_i(j), \tag{1}$$

where $P_i(j)$ denotes the predicted probability of the j-th token in the top-k candidates at step i, and k is the number of top tokens considered. High C_i corresponds to a more peaked distribution and thus greater model certainty.

Sliding-window Group Metrics. Given token confidences $\{C_i\}_{i=1}^{N}$ (Eq. 1), where N is the total number of generated tokens in the model's answer/solution, we form M overlapping windows $\mathcal{I}_j = [a_j, b_j] \subseteq \{1, \ldots, N\}$ using window size W and step S (with tail mode "full"). For each window, we compute the local group mean:

$$g_j = \frac{1}{|\mathcal{I}_j|}\sum_{i\in\mathcal{I}_j} C_i. \tag{2}$$

We report three summary metrics over $\{g_j\}_{j=1}^{M}$:

Group Overall Mean(GOM). We compute the average of all group means g_j to reflect the overall local confidence level across the entire generated answer/solution:

$$\bar{g} = \frac{1}{M}\sum_{j=1}^{M} g_j. \tag{3}$$

Group Standard Deviation(GSD). We compute the (population) standard deviation of all group means to measure the fluctuation of confidence along the trace:

$$\sigma_g = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\big(g_j - \bar{g}\big)^2}. \tag{4}$$

Group Coefficient of Variation(GCV). We compute the coefficient of variation to assess the relative fluctuation of confidence independent of its absolute level:

$$\mathrm{CV}_g = \frac{\sigma_g}{\bar{g}}. \tag{5}$$

3 Evaluation and Results

In this section, we report a comprehensive evaluation of various Vision-Language Models (VLMs) on the proposed *MathSight* benchmark. Our experiments aim to examine how visual factors influence mathematical reasoning at the university level, with particular attention to *image variation* (e.g., type and size) and *problem diversity* across different mathematical categories. We systematically analyze model performance across multiple visual conditions and reasoning domains to uncover the extent to which current models truly leverage visual information in complex mathematical problem solving.

3.1 Evaluated Models

We evaluate a diverse set of state-of-the-art Vision-Language Models (VLMs) that represent the current frontier of multimodal reasoning from open-source and closed-source. Our evaluation includes Gemini-2.5-Pro (Comanici et al., 2025), GPT-5, the Claude family (*Claude-3.7-Sonnet*, *Claude-4-Sonnet*, and *Claude-Opus*) (Anthropic, 2025a,b), Qwen2.5-VL (7B, 3B) (Bai et al., 2025), Doubao-1.5-Vision-Pro (Guo et al., 2025), GLM-4.1V-Thinking (Hong et al., 2025), InternVL3 (78B) (Zhu et al., 2025), and Qwen3-VL-235B-A22B.

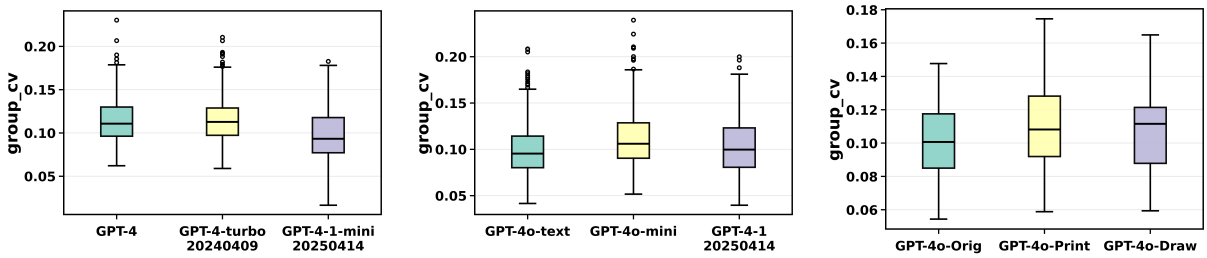


Figure 2: Box plots of group coefficient of variation (group_CV) across different models and k settings. Each box covers the 95% central region of group_CV values, the horizontal line inside the box marks the median group_CV, and the points outside indicate a small number of outliers. Left: text-only model with $k = 5$. Middle: text-only model with $k = 20$. Right: multimodal GPT-4o variants with $k = 20$.

To further disentangle the contribution of visual information, we also evaluate the text-only variant Qwen3-235B-A22B, which shares the same underlying language backbone as Qwen3-VL but operates purely on textual input. This comparison allows us to quantify how much of the reasoning capability stems from linguistic priors versus vision-grounded understanding. All models are tested in a zero-shot setting using their publicly available APIs or official implementations, ensuring reproducibility and fair comparison.

3.2 Experimental Setup

To ensure the reproducibility of our experiments, we describe the evaluation metric and the prompt templates used during model inference and evaluation in detail.

We report **accuracy (ACC)** for non-proof questions. For proof problems, we evaluate **logical consistency** of the reasoning process instead of matching a specific expression. Specifically, we use three complementary metrics—**GOM**, **GSD**, and **GCV**—to capture overlap with essential logical dependencies, step-level deviation from sound inference, and stability of consistency across the proof.

All experiments are conducted in a zero-shot setting. We obtain model responses from two sources: 1) API-based access to proprietary or cloud-hosted models, and 2) local inference using open-source models with the vLLM(Kwon et al., 2023) library. Final accuracy is computed by comparing model outputs with ground truth using mathematical equivalence, depending on the task type. We will release our benchmark on GitHub in the near future.

3.3 Overall Results

Table 2 presents the overall performance of representative vision-language models (VLMs) on the MathSight benchmark under three visual conditions: *original*, *hand-drawn*, and *photo-captured* image variants. Across both closed-source and open-source model, we observe a consistent trend: although accuracy slightly varies across image variants, the difference remain marginal, suggesting that visual appearance has only a limited impact on university-level mathematical reasoning. This finding indicates that current VLMs may tend to rely predominantly on textual cues or memorized symbolic patterns rather than genuinely integrating visual understanding into the reasoning process.

Interestingly, several models exhibit slightly higher accuracy on the *hand-drawn* and *photo-captured* variants, indicating that imperfect or naturalistic visuals do not necessarily impede comprehension. Instead, such inputs may encourage models to depend more on symbolic and relational reasoning rather than superficial visual pattern matching. This finding further supports the hypothesis that contemporary VLMs predominantly rely on textual abstraction rather than genuinely grounded visual understanding.

Moreover, MathSight, a university-level benchmark for assessing advanced multimodal mathematical reasoning, remains highly challenging. Even the state-of-the-art model, GPT-5, achieves only 45.39% accuracy, indicating a considerable gap from human-level competence. Although closed-source models such as Gemini and Claude generally outperform open-source counterparts, the performance gap is notably small. In fact, we find that tasks correctly solved by one model are often solved by most others, whereas those failed by one tend to be universally difficult across models. This

Table 2: Overall evaluation results of Vision-Language Models (VLMs) on the **MathSight** benchmark.

Model	V.Orig	V.Draw	V.Photo
Closed-source Models			
Claude-3.7-sonnet-20250219	36.80	39.64	40.39
Claude-3.7-sonnet-20250219-thinking	34.80	37.52	41.15
Claude-sonnet-4-20250514	36.16	37.67	40.70
Claude-sonnet-4-20250514-thinking	36.61	**41.60**	**41.45**
Claude-opus-4-20250514	36.61	37.97	39.79
Claude-opus-4-20250514-thinking	35.55	41.45	42.36
Gemini-2.5-pro-preview-06-05	37.07	38.73	40.85
GPT-4o(2024/05/13)	36.91	36.46	37.52
GPT-5	**45.39**	41.50	40.85
Open-source Models			
InternVL3-78B	39.64	39.49	41.30
GLM-4.1V-9B-Thinking	35.85	37.37	38.58
Qwen2.5-vl-7B	34.64	34.95	36.31
Qwen2.5-vl-3B	33.74	36.61	36.61
Qwen3-VL-235B-A22B	40.85	38.73	39.33

pattern suggests that current VLMs occupy a narrow reasoning space and exhibit highly correlated failure modes, revealing a lack of complementary multimodal capabilities across architectures.

3.4 Results on Various Categories

To further analyze how visual information interacts with different forms of reasoning, we evaluate VLMs' performance on the *original-variant* setting across various mathematical categories, including Calculus, Algebra, Applied Mathematics, Probability & Statistics, Analysis, and Discrete Mathematics. As shown in Table 3, we observe substantial variation across categories, revealing that visual dependence is not uniform but highly task-specific. Models generally achieve the highest accuracy on *algebraic* and *probabilistic* problems, where reasoning can often proceed from symbolic or textual cues alone. In contrast, performance on *calculus* and *analysis* mathematics is consistently lower, indicating that tasks requiring conceptual abstraction or geometric interpretation remain challenging for current VLMs. Interestingly, *applied mathematics* and *probability* problems—where visual representations are closely aligned with textual descriptions—exhibit comparatively higher accuracy, suggesting that models can partially exploit structured visual layouts when visual–textual alignment is strong. However, this advantage likely stems from surface-level correlation matching rather than genuine visual reasoning, underscoring that current VLMs rely heavily on linguistic priors rather than grounded visual understanding.

3.5 Results on Different-size Images

To investigate the sensitivity of Vision-Language Models (VLMs) to visual scale and resolution, we further evaluate their performance under four controlled visual configurations: *large* and *small* versions of both *hand-drawn* and *photo-captured* figures. The results are presented in Figure 3. Across all models, we observe that **variations in image size have minimal influence on overall accuracy**. For most VLMs, the difference between large and small images within the same visual type remains within a narrow range of 1–2 percentage points, far smaller than the variance typically observed across problem categories or modalities. This stability suggests that current VLMs rely little on fine-grained visual detail when solving mathematical problems. Instead, their reasoning appears to be dominated by textual understanding and symbolic pattern recognition extracted from the accompanying problem statement.

Interestingly, for some models such as **Gemini 2.5 Pro** and **Claude 3.7 Sonnet**, performance with smaller hand-drawn or photo images is even slightly higher than with larger ones. This counterintuitive behavior indicates that resizing visual inputs does not degrade, and may occasionally regularize, model predictions—possibly because smaller images implicitly downweight the visual embedding relative to the text, thereby reducing visual noise.

Taken together, these results suggest that current VLMs display a high degree of **scale invariance** yet a low degree of **visual reliance**. In other words, altering the scale of the visual input scarcely impacts reasoning performance, implying that these models engage only superficially with visual information and instead depend predominantly on linguistic priors.

3.6 Text-Only vs. Multimodal Reasoning on VLMs

To directly assess the contribution of visual input, we evaluate **Qwen3-VL-235B-A22B** under two settings: a *text-only* setting and a *multimodal* setting. In the text-only setting, Qwen3-VL only receives the full problem statement but no accompanying visual content, allowing us to isolate whether the visual modality provides complementary grounding or merely redundant information. In the multimodal configuration, we further examine three

Table 3: Evaluation results of Vision-Language Models (VLMs) under different subjects. (Note that subject abbreviations indicate: Applied. = Applied Math, Pro.&Sta. = Probability & Statistics, Discrete. = Discrete Mathematics). The **bold** and underline numbers represent the best and the second best results, respectively.

Model	All 661	Calculus 543	Algebra 7	Applied. 78	Pro.&Sta. 21	Analysis 6	Discrete. 6
Closed-source Models							
Claude-3.7-sonnet-20250219	36.76	29.10	71.43	73.08	76.19	50.00	66.67
Claude-3.7-sonnet-20250219-thinking	34.80	27.07	71.43	69.23	80.95	50.00	66.67
Claude-sonnet-4-20250514	36.16	28.55	71.43	71.79	80.95	33.33	66.67
Claude-sonnet-4-20250514-thinking	37.97	30.57	71.43	70.51	85.71	50.00	66.67
Claude-opus-4-20250514	37.52	29.65	71.43	74.36	85.71	33.33	66.67
Claude-opus-4-20250514-thinking	35.55	28.36	71.43	69.23	71.43	50.00	66.67
Gemini-2.5-pro-preview-06-05	37.07	28.73	85.71	73.08	**90.48**	50.00	66.67
GPT-4o(2024/05/13)	38.43	31.12	85.71	69.23	80.95	66.67	66.67
GPT-5	**45.39**	37.94	85.71	79.49	85.71	66.67	66.67
Open-source Models							
InternVL3-78B	**39.64**	**32.78**	**71.43**	**70.51**	76.19	**66.67**	66.67
GLM-4.1V-9B-Thinking	35.85	28.36	71.43	70.51	76.19	50.00	66.67
Qwen2.5-vl-7B	34.64	27.44	71.43	66.67	**80.95**	33.33	66.67
Qwen2.5-vl-3B	33.74	25.97	71.43	69.23	76.19	50.00	66.67
Qwen3-VL-235B-A22B	40.85	33.70	85.71	71.79	85.71	50.00	66.67

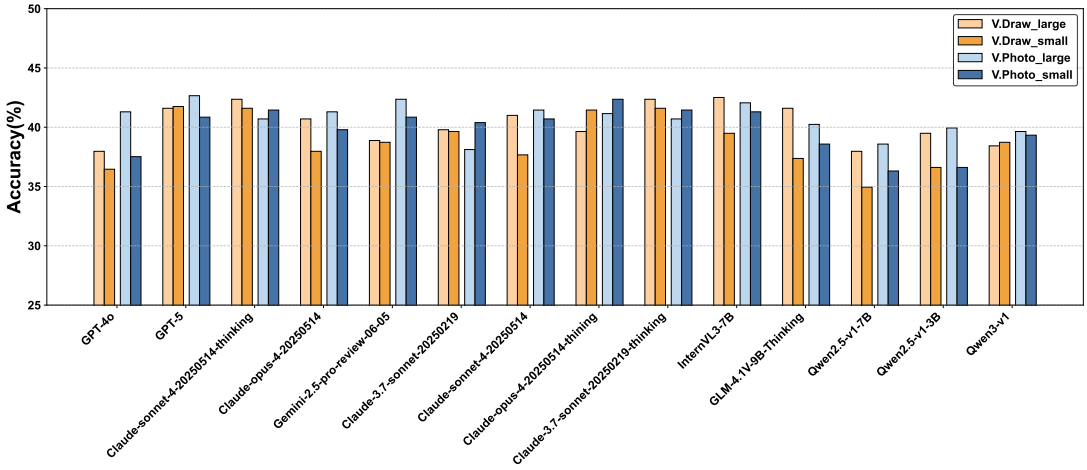


Figure 3: Evaluation results of Vision-Language Models (VLMs) on images of different sizes.

types of visual inputs such as the *original*, *hand-drawn*, and *photo-captured* variants to analyze how different visual representations influence the performance of university-level mathematical reasoning.

As shown in Figure 4, we observe that **removing the image input leads to a substantial performance improvement**, boosting accuracy from 40.85% to 50.53% and even outperforming the state-of-the-art GPT-5 (45.39%). This counterintuitive result suggests that current VLMs may not truly engage in visual reasoning but rather treat visual inputs as noisy or even misleading signals when solving mathematically complex problems. The improvement in the text-only setting indicates that current VLMs may rely heavily on linguistic priors and symbolic patterns learned during pre-training, rather than deriving meaningful information from visual input. This finding highlights a fundamental limitation of current VLMs' architectures: the presence of an image does not guarantee vision-grounded understanding, especially in university-level mathematics.

In addition, we extend the text-only evaluation to the **Qwen3-LM** model, which shares the same language backbone as Qwen3-VL but lacks any visual processing capability. Interestingly, its accuracy drops significantly to **24.21%**, far below both the multimodal and text-only variants of Qwen3-VL. This sharp contrast indicates that the superior text-only performance of Qwen3-VL does not stem merely from its linguistic backbone but from its training as a multimodal model, which appears to internalize certain structural priors that generalize even without visual input.

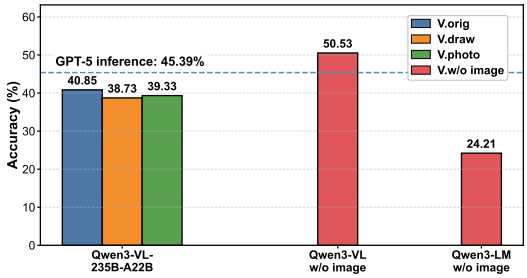


Figure 4: The result comparison between text-only and different multimodal inputs.

3.7 Further Analysis

To further investigate how visual variations influence reasoning stability, we analyze the consistency of model predictions across the three image variants (*original*, *hand-drawn*, and *photo-captured*). Table 4 reports the distribution of answer combinations for GPT-5 and Qwen3-VL-235B-A22B.

We observe that the performance consistency across different image variants is extremely high for both GPT-5 and Qwen3-VL. More than **80%** of all cases remain fully consistent, either all correct (*A✓B✓C✓*) or all incorrect (*A✗B✗C✗*), while only about **18%** show any variation across visual versions. Such strong consistency implies that model predictions are largely insensitive to changes in visual form, suggesting that performance is driven primarily by linguistic or symbolic reasoning rather than genuine visual grounding. In other words, whether the accompanying image is original, hand-drawn, or captured in a photo has little influence on the outcome: the models effectively "read" but do not "see". This finding underscores that current VLMs treat visual input as an auxiliary cue rather than an integral reasoning component, revealing a fundamental limitation in their capacity for truly vision-conditioned mathematical understanding.

4 Related Work

In order to study the effects of mathematical diagrams on problem-solving, various multimodal benchmarks are proposed (Chen et al., 2023; Lu et al., 2023; Gupta et al., 2024; Wang et al., 2023; Zhang et al., 2024a; Wang et al., 2024; Yue et al., 2024; Liu et al., 2024b; Zou et al., 2024; Zhou et al., 2024; Chernyshev et al., 2024). Recent multimodal mathematical benchmarks are not only focus on geometric figures (Zhang et al., 2024a) but also broaden to wider range, such as venn diagrams, spatially-related layouts (Gupta et al., 2024).

Table 4: The percentage of different answer combinations on three image versions, GPT-5 and Qwen3-VL-235B-A22B as examples. Note that A denotes *original variant*, B represents *photo-captured variant*, and C corresponds to *hand-drawn variant*.

Answer Comparison	GPT-5	Qwen3-VL-235B-A22B
A✓B✓C✓	34.19%	30.56%
A✓B✓C✗	3.48%	3.48%
A✓B✗C✓	2.87%	2.87%
A✓B✗C✗	4.84%	3.93%
A✗B✓C✓	1.36%	2.42%
A✗B✓C✗	2.72%	2.87%
A✗B✗C✓	2.42%	2.87%
A✗B✗C✗	48.11%	50.98%

MMMU (Yue et al., 2024) offers multi-discipline reasoning, and its limited math subset consists of graph functions and mathematical notation images. U-MATH (Chernyshev et al., 2024) is composed of 1100 enclosed open-ended university-level problems and an extra dataset to testify the judging abilities of LLMs. Similar to UGMathBench (Xu et al., 2025), DYNAMATH (Zou et al., 2024) is also a dynamic benchmark but with visual information, which includes 501 high-quality, multi-topic seed questions. Though these benchmarks cover multimodal input, they lack detailed exploration on the effectiveness of the image attributes. In MathSight, we not only collected diverse-category problems but also augmented the image into different variants, which is an outstanding superiority for visual understanding research.

5 Conclusion

We introduced MathSight, a university-level benchmark designed to isolate and measure the value of visual information in multimodal mathematical reasoning. Each problem is paired with multiple visual variants (original, hand-drawn, photo) and a text-only condition, enabling controlled comparisons. We also provide a difficulty-matched, text-only benchmark to calibrate language-model reasoning. Across state-of-the-art VLMs, our results are clear: as problems get harder, the benefit of images shrinks. Performance differences among visual variants are not statistically significant, and strong models often match or surpass multimodal results using only text—evidence that current VLMs lean heavily on textual priors rather than genuinely vision-grounded reasoning.

6 Limitations

While MathSight offers a focused lens on vision-grounded mathematical reasoning, it has several limitations. **Model coverage:** Our experiments primarily evaluate the latest Qwen-series VLMs, with limited exploration of other model families. Broader cross-family evaluation (e.g., OpenAI, Google, Anthropic, Meta, Mistral) is needed to validate the generality of our findings. **Dataset scale:** The multimodal portion of the benchmark is modest in size. Larger, more diverse visual sets—covering varied diagram styles, noise conditions, and real-world capture artifacts—would improve statistical power and ecological validity. **Benchmark breadth:** Our conclusions should be tested across a wider range of multimodal math benchmarks and related tasks to assess robustness, transferability, and potential domain-specific effects. **Task scope:** MathSight focuses on university-level problems; extending to lower grades, professional domains (e.g., engineering drawings), and step-by-step solution formats may reveal different dependencies on visual information. We view these as practical next steps to strengthen external validity and to better separate linguistic priors from genuinely vision-grounded reasoning.

References

Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards Interpretable Math Word Problem Solving with Operation-Based Formalisms. *Computing Research Repository*, arXiv:1905.13319.

Anthropic. 2025a. Claude 3.7 sonnet system card. https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf.

Anthropic. 2025b. Introducing claude 4. https://www.anthropic.com/news/claude-4.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. Theoremqa: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7901.

Konstantin Chernyshev, Vitaliy Polshkov, Ekaterina Artemova, Alex Myasnikov, Vlad Stepanov, Alexei Miasnikov, and Sergei Tilga. 2024. U-math: A University-Level Benchmark for Evaluating Mathematical Skills in LLMs. *Computing Research Repository*, arXiv:2412.03205.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Computing Research Repository*, arXiv:2110.14168.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *Computing Research Repository*, arXiv:2507.06261.

Jingxuan Fan, Sarah Martinson, Erik Y. Wang, Kaylie Hausknecht, Jonah Brenner, Danxian Liu, Nianli Peng, Corey Wang, and Michael P. Brenner. 2024. Hardmath: A Benchmark Dataset for Challenging Problems in Applied Mathematics. *Computing Research Repository*, arXiv:2410.09988.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. 2025. Deep think with confidence. *Preprint*, arXiv:2508.15260.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, Jingji Chen, Jingjia Huang, Kang Lei, Liping Yuan, Lishu Luo, Pengfei Liu, Qinghao Ye, Rui Qian, Shen Yan, and 178 others. 2025. Seed1.5-vl technical report. *arXiv preprint arXiv:2505.07062*.

Himanshu Gupta, Shreyas Verma, Ujjwala Anantheswaran, Kevin Scaria, Mihir Parmar, Swaroop Mishra, and Chitta Baral. 2024. Polymath: A Challenging Multi-Modal Mathematical Reasoning Benchmark. *Computing Research Repository*, arXiv:2410.14702.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *Computing Research Repository*, arXiv:2009.03300.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *Computing Research Repository*, arXiv:2103.03874.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, and 1 others. 2025. Glm-4.1 v-thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning. *Computing Research Repository*, arXiv:2507.01006.

Xuhan Huang, Qingning Shen, Yan Hu, Anningzhe Gao, and Benyou Wang. 2024. Mamo: A Mathematical Modeling Benchmark with Solvers. *Computing Research Repository*, arXiv:2405.13144.

Zhewei Kang, Xuandong Zhao, and Dawn Song. 2025. Scalable best-of-n selection for large language models via self-certainty. *Preprint*, arXiv:2502.18581.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. *Computing Research Repository*, arXiv:2309.06180.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. 2024. Gsm-plus: A Comprehensive Benchmark for Evaluating the Robustness of LLMs as Mathematical Problem Solvers. *Computing Research Repository*, arXiv:2402.19255.

Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024a. Mathbench: Evaluating the Theory and Application Proficiency of LLMs with a Hierarchical Mathematics Benchmark. *Computing Research Repository*, arXiv:2405.12209.

Xiao Liu, Zirui Wu, Xueqing Wu, Pan Lu, Kai-Wei Chang, and Yansong Feng. 2024b. Are llms capable of data-based statistical and causal reasoning? Benchmarking Advanced Quantitative Reasoning with Data. *Computing Research Repository*, arXiv:2402.17644.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and 1 others. 2022. Lila: A Unified Benchmark for Mathematical Reasoning. *Computing Research Repository*, arXiv:2210.17517.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *Computing Research Repository*, arXiv:2402.14804.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models. *Computing Research Repository*, arXiv:2307.10635.

Xin Xu, Jiaxin Zhang, Tianhao Chen, Zitong Chao, Jishan Hu, and Can Yang. 2025. Ugmathbench: A Diverse and Dynamic Benchmark for Undergraduate-Level Mathematical Reasoning with Large Language Models. *Computing Research Repository*, arXiv:2501.13766.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of CVPR*.

Jiaxin Zhang, Zhongzhi Li, Mingliang Zhang, Fei Yin, Chenglin Liu, and Yashar Moshfeghi. 2024a. Geoeval: Benchmark for Evaluating LLMs and Multi-Modal Models on Geometry Problem-Solving. *Computing Research Repository*, arXiv:2402.10104.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024b. 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. Springer.

Zihao Zhou, Shudong Liu, Maizhen Ning, Wei Liu, Jindong Wang, Derek F. Wong, Xiaowei Huang, Qiufeng Wang, and Kaizhu Huang. 2024. Is your model really a good math reasoner? Evaluating Mathematical Reasoning with Checklist. *Computing Research Repository*, arXiv:2407.08733.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *Computing Research Repository*, arXiv:2504.10479.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. 2024. Dynamath: A Dynamic Visual Benchmark for Evaluating Mathematical Reasoning Robustness of Vision Language Models. *Computing Research Repository*, arXiv:2411.00836.

Appendix —— MathSight

A Comparison of Existing Benchmarks in Text-only and Multimodal

Table 5: Existing Benchmarks, where "Vi.Var" indicates that visual input has different variants and "Pro.Q" stands for proving questions, i.e., questions that cannot be simply verified against a standard answer.

Benchmark	Venue	Uni. Level	Other Level	Pure-text	Multimodal	Grad. Level	Pro.Q	Vi.Var
MathQA (Amini et al., 2019)	ACL'19	0	37,200	37,200	0	✗	✗	✗
$MMLU_{math}$ (Hendrycks et al., 2020)	ICLR'21	200	648	848	0	✗	✗	✗
GSM8K (Cobbe et al., 2021)	arXiv'21	0	1,319	1,319	0	✗	✗	✗
Math (Hendrycks et al., 2021)	NeurIPS'21	0	5,000	5,000	0	✗	✗	✗
Lila (Mishra et al., 2022)	EMNLP'22	86,742	47,073	133,815	0	✗	✗	✗
GSM-Plus (Li et al., 2024)	ACL'24	0	10,552	10,552	0	✗	✗	✗
MathBench (Liu et al., 2024a)	ACL Findings'24	932	2,777	3,709	0	✗	✗	✗
Mamo (Huang et al., 2024)	arXiv'24	557	652	1,209	0	✗	✗	✗
PolyMATH (Gupta et al., 2024)	arXiv'24	6,750	2,250	9000	0	✗	✗	✗
HARDMath (Fan et al., 2024)	ICLR'25	1,466	0	1,466	0	✓	✗	✗
UGMathBench (Xu et al., 2025)	ICLR'25	5,062	0	5,062	0	✗	✗	✗
TheoremQA (Chen et al., 2023)	EMNLP'23	800	0	749	51	✗	✗	✗
MathVista (Lu et al., 2023)	ICLR'24	661	5,480	0	6,141	✗	✗	✗
Scibench (Wang et al., 2023)	ICML'24	869	0	692	177	✗	✗	✗
QRData (Liu et al., 2024b)	ACL Findings'24	411	0	0	411	✗	✗	✗
MATH-Vision (Wang et al., 2024)	NeurIPS'24	0	3,040	0	3,040	✗	✗	✗
$MMMU_{math}$(Yue et al., 2024)	CVPR'24	540	0	0	540	✗	✗	✗
U-Math (Chernyshev et al., 2024)	arXiv'24	1,100	0	880	220	✗	✗	✗
Dynamath (Zou et al., 2024)	ICLR'25	1,610	3,400	310	4,700	✗	✗	✗
MathCheck (Zhou et al., 2024)	ICLR'25	0	4,536	3,096	1,440	✗	✗	✗
MathSight (our work)	NA	2048	0	1,387	661	✓	✓	✓

B Dataset Composition

B.1 Overview of MathSight

Figure 5 is overview (Categories and examples) of MathSight.

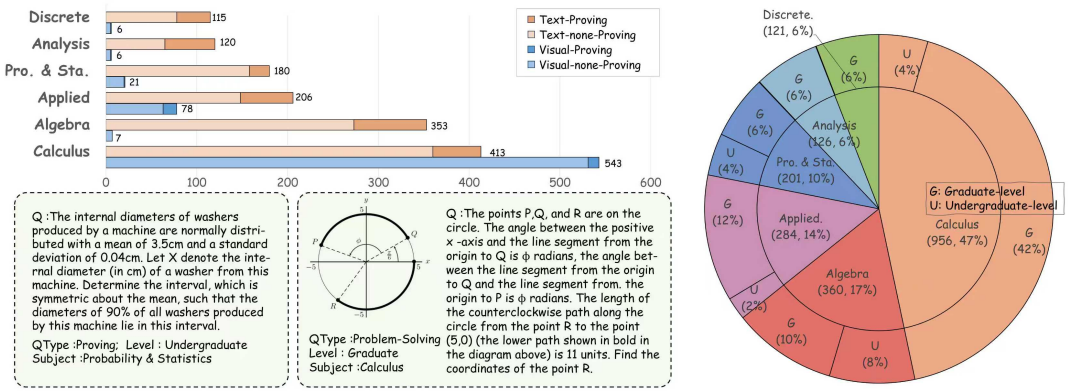


Figure 5: Overview (Categories and examples) of MathSight.

B.2 The Pipeline of MathSight Construction

Figure 6 depicts the construction pipeline of MathSight.

B.3 Mathematics Problem Extraction Task Processing Template (Reusable)

A reusable extraction strategy was designed to guide the model, ensuring consistently well-formatted outputs. A detailed and reproducible version of this extraction strategy is presented below.

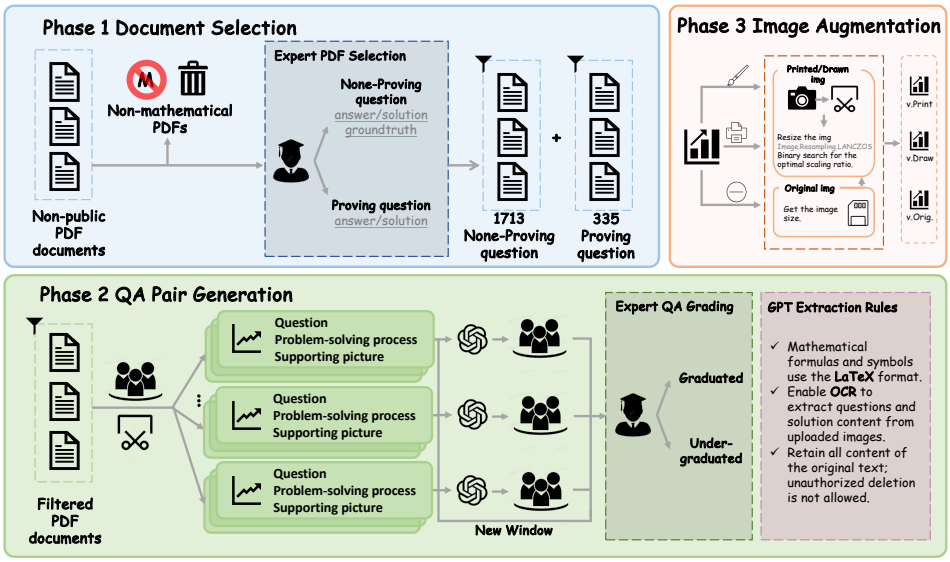


Figure 6: The construction pipeline of MathSight.

1. General Processing Requirements

1. **Output Format.** Output format should be JSONL. Each line must contain the following fields:
   - `question`
   - `answer/solution`
   - `ground_truth`
   - `image_path`
2. **Preservation of Original Text.** Both the problem statement and the solution must be fully retained in their original form (including derivations, symbols, and formulas). Simplification, summarization, or editing is not allowed.
3. **Sub-question Handling.**
   - If the problem contains sub-questions (a)(b)(c)..., all sub-questions must supplement the full main problem statement.
   - Remove markers such as (a)/(b) and maintain continuous language.
   - Each sub-question must keep the entire main problem statement completely intact without any omission.
4. **Enable OCR.** Enable image recognition (OCR) to read problem statements and solutions from uploaded images or PDFs.
5. **Correct Understanding of `ground_truth`.** For non-proof problems with a final computed result (e.g., equals some expression), the `ground_truth` field must record this final result. If there is no final result, leave the `ground_truth` field blank. If the question asks for the determined answers of multiple parameters, `ground_truth` should list each parameter's final determined answer. For True/False questions, `ground_truth` should be marked as `true` or `false` accordingly.
6. **Mathematical Notation in LaTeX.** All mathematical symbols and letters must follow LaTeX math-reasoning standard formatting. For example, subscripts in LaTeX should always be written as a_i or $a_{...}$.
7. **Mandatory LaTeX for Formulas and Symbols.** Always use LaTeX formatting for mathematical formulas and symbols; otherwise, errors may occur.

2. Skipping and Exception Handling Rules

1. **Skipping Problems Without Solutions.** If a problem has no solution (the `solution/answer` field is empty), skip that problem.
2. **Recording Skipped Problems.** Record the following information for each skipped problem:
   - File name
   - Problem number
   - Reason for skipping
3. **If Processing Stops Midway.** Indicate:
   - Which file and which problem (statement or page) it stopped at
4. **Handling Images in Answers or Requests.** For answers or problems requiring images, set `image_path` to the image's file name. If no image exists, leave it blank. If an image is uploaded, set that problem's `image_path` to the image's name.
5. **Preserve Full Original Content.** All original content must be preserved without unauthorized deletion. Words may be added only to maintain fluency.

3–6. Additional Processing Instructions

1. **Pre-processing Instructions.** Explain how to paginate the processing and how to maintain high-quality results. Provide a table indicating whether `ground_truth` and `image_path` are used. Indicate whether multiple sub-questions in one problem can be recognized and split.
2. **During Processing.** Process problems in groups of three. After each group is processed, export the results as a JSONL file and make it downloadable.
3. **Exporting.** From the first problem onward, export as JSONL. You may export in segments to ensure quality.
4. **Additional Requirements for Specific Problems.** (Additional requirements to be filled as needed.)

C Prompt

C.1 Generation Prompt

To generate model responses, we use the following zero-shot prompt:

Zero-shot Prompt

You are a university-level mathematics instructor. Given the following question, provide a complete and rigorous solution. The final answer must be clearly marked as ‘groundtruth‘.
If the question is a proof problem, provide the full proof in the solution section, but leave the groundtruth field empty.
Please strictly follow this output format (use English field names exactly as shown):

question:
<Original question text>
answer/solution:
<Your detailed solution>
groundtruth:
<Final answer or leave empty if proof>

C.2 Evaluation Prompt

To assess whether the model’s answer is correct, we use the following evaluation prompts. Specifically, we have adopted different prompts for proof questions and non-proof questions respectively:

Evaluation Prompt for Proof Questions

You are a mathematics instructor. Determine whether the following student’s proof is logically valid and correctly proves the statement. Ignore formatting, focus only on mathematical logic.

Question: {question}
Reference Proof: {ref_solution}
Student’s Proof: {student_solution}
Return only "Yes" if the proof is correct or "No" otherwise.

Evaluation Prompt for Non-Proof Questions

You are a mathematics evaluator. Check whether the student’s final answer and ground_truth are mathematically equivalent to the reference answer and reference ground_truth. Ignore formatting differences.

Question: {question}
Reference Final Answer: {ref_answer}
Student’s Final Answer: {student_answer}

Return only “Yes” if the answer is correct or “No” otherwise.

D Subcategories and Keywords of the *Calculus* Primary Category in MathSight.

As shown in Table 3, the multimodal portion of the MathSight dataset contains 543 problems under the *Calculus* primary category. This coarse categorization may lead to a perception of imbalance within the multimodal dataset. To address this, we further subdivide *Calculus* into secondary categories according to the chapter organization of *The Princeton Companion to Calculus*. This refined classification illustrates the diversity of subjects covered and highlights that several key topics are represented by a notably large number of problems.

Subcategory	Keyword (	Count
Functions and Graphs	function	309
	graph	74
	inverse	25
	trigonometric	19
	sin	14
	cos	11
	tan	8
Limits and Continuity	limit	23
	limit definition	12
	approach	8
	continuity	4
Derivatives and Rules	derivative	20
	slope	13
	rate of change	7
	tangent	5
Applications of Derivatives and Optimization	optimization	26
	max	14
	min	12
	motion	10
	velocity	7
	acceleration	5
Exponential and Logarithmic Functions	exponential	22
	logarithm	18
	ln	12
Integrals and Fundamental Theorem of Calculus	integral	36
	area	21
	area under	10
Integration Techniques	substitution	7
	integration by parts	6
Series and Taylor Expansion	series	13
	taylor	7
Differential Equations	differential equation	12
	growth	4
	decay	3

Table 6: Breakdown of Calculus Problems by Subcategory and Keyword. Each problem is uniquely assigned to one keyword. **Total:** 34 keywords, 956 problems.

E Classification Criterion of Undergraduate- and Graduate-level Problems

In this section, we present our self-defined standard for distinguishing between undergraduate- and graduate-level problems. This standard was developed based on our analysis of the dataset and aims to provide a consistent and reproducible classification framework for difficulty labeling. Specifically, the classification is based on three dimensions: curriculum mapping, knowledge depth, and classification principles.

E.1 Curriculum Mapping

The curriculum mapping provides an objective basis for distinguishing between undergraduate and graduate levels. Higher education mathematics courses follow a clear progression in content scope and theoretical depth.

Undergraduate Level *Typical Courses*: Linear Algebra, Single/Multivariable Calculus, Probability & Statistics, Ordinary Differential Equations, Introductory Numerical Analysis. *Characteristics*: Course content focuses on fundamental theorems and formula applications. Problems are usually within the scope of standard textbook exercises, with relatively simple mathematical notation and limited derivations.

Graduate Level *Typical Courses*: Real Analysis, Complex Analysis, Functional Analysis, Partial Differential Equations, Advanced Probability Theory & Stochastic Processes, Topology & Geometry, Optimization Theory, Numerical PDE. *Characteristics*: Course content emphasizes the rigor of theoretical systems and the understanding of abstract concepts. Problems often originate from graduate-level textbooks, research papers, or comprehensive exams, with more complex notation systems and multi-layered derivations.

E.2 Knowledge Depth

Undergraduate Level: Problems typically involve direct applications of core undergraduate courses, with known formulas directly substituted or solved with a single-step derivation. Reasoning chains are short (1–3 steps), and problems are often closed-ended, with given conditions directly leading to the answer. **Graduate Level**: Problems often require formula derivation, proof of conclusions, or integration of knowledge from multiple disciplines. Reasoning chains are long (more than 3 steps), and problems are often open-ended, with multiple possible solution paths.

E.3 Classification Principles

- **Disciplinary Breadth Principle**: If a problem spans two or more advanced mathematical fields (e.g., Real Analysis and Probability, Algebra and Topology), it is likely classified as graduate level.
- **Knowledge Depth Principle**: If a problem involves mathematical concepts not covered at the undergraduate level, it is classified as graduate level.
- **Solution Requirement Principle**: If a problem requires proof of a general proposition rather than computing a specific example, it is likely classified as graduate level.

F Evaluated Large Models

We provide more detailed information about the evaluated large models in Table 7.

G Evaluation Results on Different Question Types

We further analyzed the accuracies of different models on proving and non-proving problems. As shown in Figure 7, our problem set contains 380 undergraduate-level problems, among which 50 are proving problems. The graduate-level set contains 1668 problems, with 285 of them being proving problems. Therefore, we can calculate the proportion: proving problems account for 13.16% of the undergraduate-level problems, while they account for 17.09% of the graduate-level problems. This difference is one of the reasons why, in Figure 8, the accuracy of undergraduate-level problems is lower than that of graduate-level problems.

Table 7: Evaluated Large Models.

LLM	Size	Organization	Multi-modal	Open-source
Qwen2.5-Math	72B, 7B	Alibaba	✗	✓
Qwen2.5-vl	7B, 3B	Alibaba	✓	✓
Qwen3	32B, 14B, 4B, 8B	Alibaba	✗	✓
InternVL3	78B	OpenGVLab	✓	✓
DeepSeekV3 (2024/12/26)	671B	DeepSeek AI	✗	✓
DeepSeekR1 (2025/1)	671B	DeepSeek AI	✗	✓
GLM-Z1-Air	32B	Zhipu AI & Tsinghua University	✗	✓
GLM-4.1V-Thinking	9B	Zhipu AI & Tsinghua University	✓	✓
Kimi-k2-0711-preview	1T	Moonshot AI	✗	✓
Claude-3-7-sonnet-20250219	-	Anthropic	✓	✗
Claude-3-7-sonnet-20250219-thinking	-	Anthropic	✓	✗
Claude-sonnet-4-20250514	-	Anthropic	✓	✗
Claude-sonnet-4-20250514-thinking	-	Anthropic	✓	✗
Claude-opus-4-20250514	-	Anthropic	✓	✗
Claude-opus-4-20250514-thinking	-	Anthropic	✓	✗
Doubao-seed-1.6-250615	-	ByteDance	✗	✗
Doubao-seed-1.6-thinking-250615	-	ByteDance	✗	✗
Doubao-1.5-thinking-vision-pro-250428	-	ByteDance	✓	✗
Gemini-2.5-pro-preview-06-05	-	Google DeepMind	✓	✗
Gemini-2.5-pro-preview-06-05-thinking	-	Google DeepMind	✓	✗
GLM-4-Plus	-	Zhipu AI & Tsinghua University	✗	✗
GPT-4	-	OpenAI	✗	✗
GPT-4-turbo (2024/4/9)	-	OpenAI	✗	✗
GPT-4.1 (2025/04/14)	-	OpenAI	✗	✗
GPT-4.1-mini (2025/04/14)	-	OpenAI	✗	✗
GPT-4o-mini	-	OpenAI	✗	✗
GPT-o4-mini (2025/04/16)	-	OpenAI	✗	✗
GPT-4o (2024/5/13)	-	OpenAI	✓	✗

In addition, as shown in Figure 9, we calculated two types of accuracies for each model in the problem set: (1) the proportion of correctly solved proving problems among all proving problems, which we call *Right-proving accuracy*; (2) the proportion of correctly solved non-proving problems among all non-proving problems, which we call *Right-non-proving accuracy*.

For example, for **Doubao-seed-1.6-250615**, the left bar represents the *Right-proving accuracy* at 91.15%, while the right bar represents the *Right-non-proving accuracy* at 33.09%. As shown in the left panel of Figure 9, large language models on the textual problem set clearly exhibit a higher *Right-proving accuracy* than *Right-non-proving accuracy*, which explains why in Figure 8 the accuracy for undergraduate-level problems is lower than for graduate-level problems.

Meanwhile, as shown in the right panel of Figure 9, for multi-modal models the *Right-proving accuracy* is slightly lower than the *Right-non-proving accuracy*. However, since the multi-modal set contains only 30 proving problems out of 661 problems in total, the proportion is very small. As a result, the difference between *Right-proving accuracy* and *Right-non-proving accuracy* has minimal impact on the undergraduate- and graduate-level accuracies of multi-modal models shown in Figure 8.

G.1 The Comparison of Large Models on Undergraduate- and Graduate-level Data

As shown in Figure 8, we calculated the accuracies of large models on undergraduate- and graduate-level problems separately. *Counterintuitively*, the performance of large models on graduate-level problems exceeds that on undergraduate-level problems. Upon further analysis, we identified two possible reasons for this result. First, the number of undergraduate-level problems in UMMBench is noticeably smaller than that of graduate-level problems, which may lead to insufficient reflection of the models' capabilities at the undergraduate level. Second, undergraduate-level problems contains more items in *Algebra* and

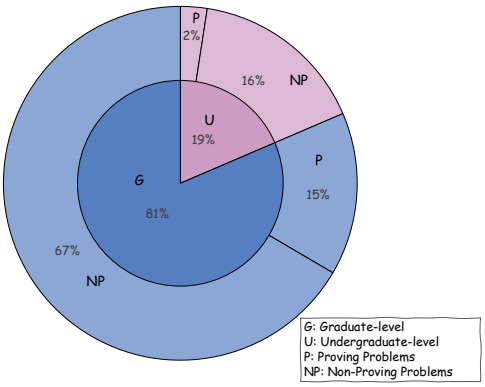


Figure 7: The proportion of different question types at the undergraduate and graduate levels.

Probability & Statistics, as shown in Figure 8. These items need multi-step computations more, which may cause large models to accumulate excessive numerical errors, ultimately leading to a drop in accuracy. These findings may also suggest that large models perform better on proving problems.

G.2 Error Analysis

We conduct error analysis for the textual model **Doubao-seed-1.6-250615** and the multimodal model **Doubao-1.5-thinking-vision-pro-250428**. From the incorrect problems of each model, we randomly sampled 100 cases for error analysis. Then we manually categorized these cases into five predefined error types (Xu et al., 2025). Specifically, *misunderstanding* refers to incorrect interpretation of the problem; *instruction following error* corresponds to cases where the solution format or required expression was not followed; *numeric calculation error* denotes mistakes in numerical computation or rounding; *express calculation error* indicates structural or symbolic inaccuracies in mathematical expressions; and *partially correct answer* describes solutions with correct intermediate reasoning but incorrect final results.

For **Doubao-seed-1.6-250615**, the most frequent error type was *partially correct answer*, indicating that the model often established an appropriate problem-solving framework but failed in later derivation or final computation. The second most common error type was *misunderstanding*, suggesting that the model sometimes misinterprets problem statements, especially in complex word problems. Errors related to *numeric calculation* and *express calculation* were relatively rare, implying that once the model correctly interprets the problem, it can usually maintain symbolic and numerical consistency.

For **Doubao-1.5-thinking-vision-pro-250428**, *misunderstanding* accounted for a significantly higher proportion of errors compared to the text-only model. This is likely due to additional challenges in interpreting visual information or aligning it with textual descriptions. While *express calculation errors* were minimal, *partially correct answer* still represented a large portion of the mistakes, indicating that even when the visual information was processed correctly, reasoning consistency issues persisted.

Overall, both models exhibited a high frequency of *partially correct answer* and *misunderstanding* errors, suggesting that improving problem comprehension and reasoning chain consistency should be priorities for future work. The text-only model tended to make fewer interpretation errors but more expression-related mistakes, whereas the multimodal model struggled more with precise problem understanding, especially in visually grounded tasks. As shown in Figure 10, partially correct answers and misunderstandings dominate the error distribution for both models, though their proportions differ considerably between the text-only and multimodal settings.

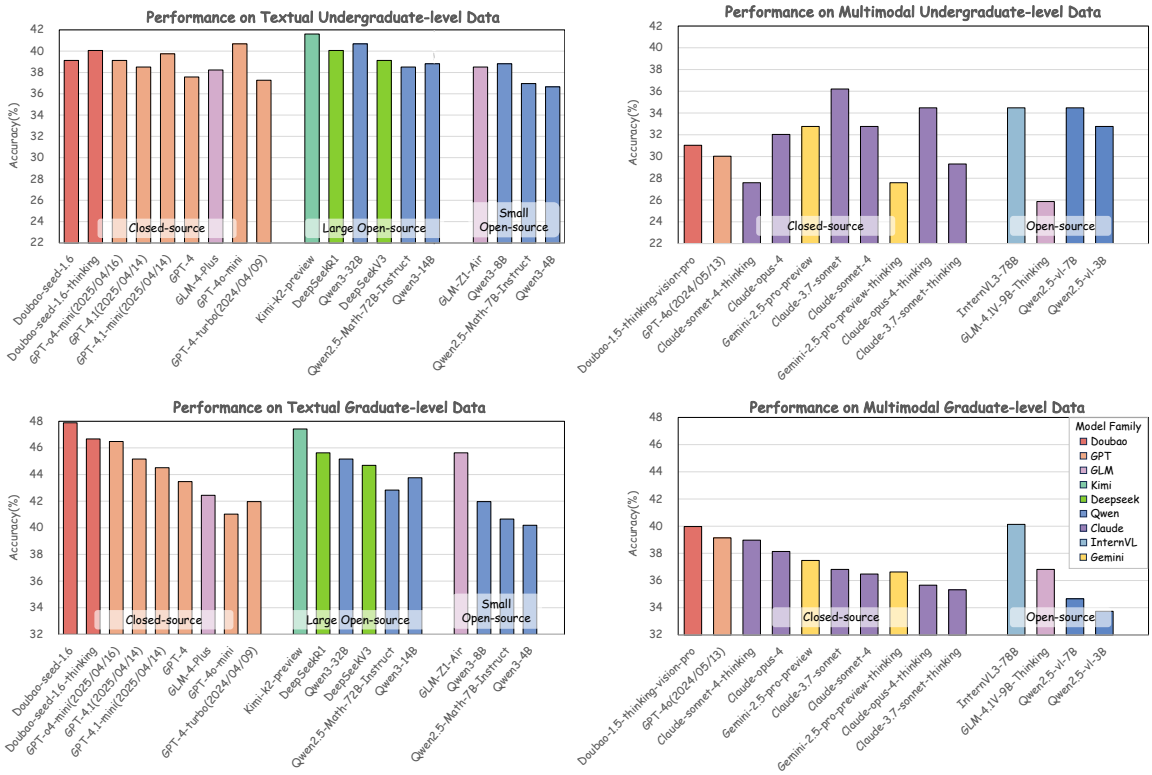


Figure 8: The proportion of different question types at the undergraduate and graduate levels.

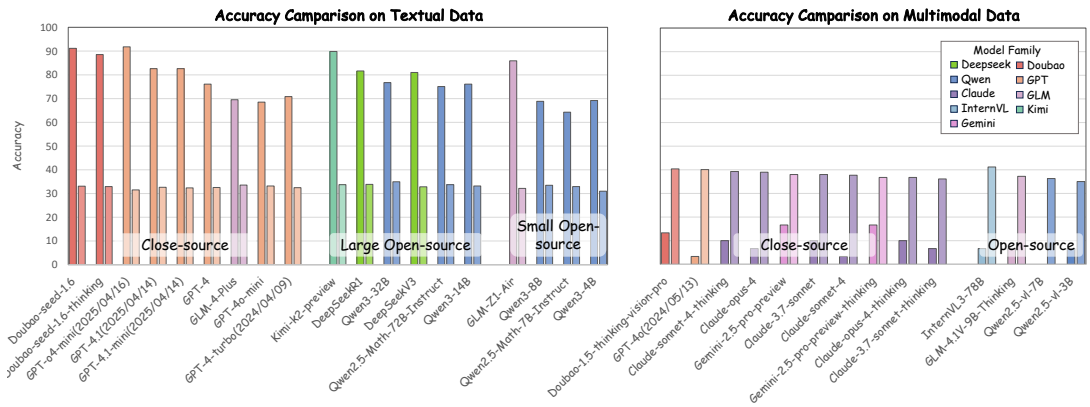


Figure 9: Accuracy comparison of proving and non-proving problems.

Table 8: Evaluation results of large models on multimodal data of UMMBench under different-sized images. The **bold** and underline numbers represent the best and second-best results of each column, respectively.

Results on Multimodal Data with Different-sized Images				
Model	$V.Draw_{large}$	$V.Draw_{small}$	$V.Photo_{large}$	$V.Photo_{small}$
Closed-source Language Models				
Claude-3.7-sonnet-20250219	39.79	39.64	38.12	40.39
Claude-3.7-sonnet-20250219-thinking	**42.36**	41.60	40.70	41.45
Claude-sonnet-4-20250514	41.00	37.67	41.45	40.70
Claude-sonnet-4-20250514-thinking	**42.36**	41.60	40.7	41.45
Claude-opus-4-20250514	40.70	37.97	41.30	39.79
Claude-opus-4-20250514-thinking	39.64	41.45	41.15	42.36
Gemini-2.5-pro-preview-06-05	38.88	38.73	42.36	40.85
GPT-4o(2024/05/13)	37.97	36.46	41.30	37.52
GPT-5	41.60	41.75	42.66	40.85
Open-source Multimodal Models				
InternVL3-78B	**42.51**	**39.49**	**42.06**	**41.30**
GLM-4.1V-9B-Thinking	41.60	37.37	40.24	38.58
Qwen2.5-vl-7B	37.97	34.95	38.58	36.31
Qwen2.5-vl-3B	39.49	36.61	39.94	36.61
Qwen3-vl	38.43	38.73	39.64	39.33

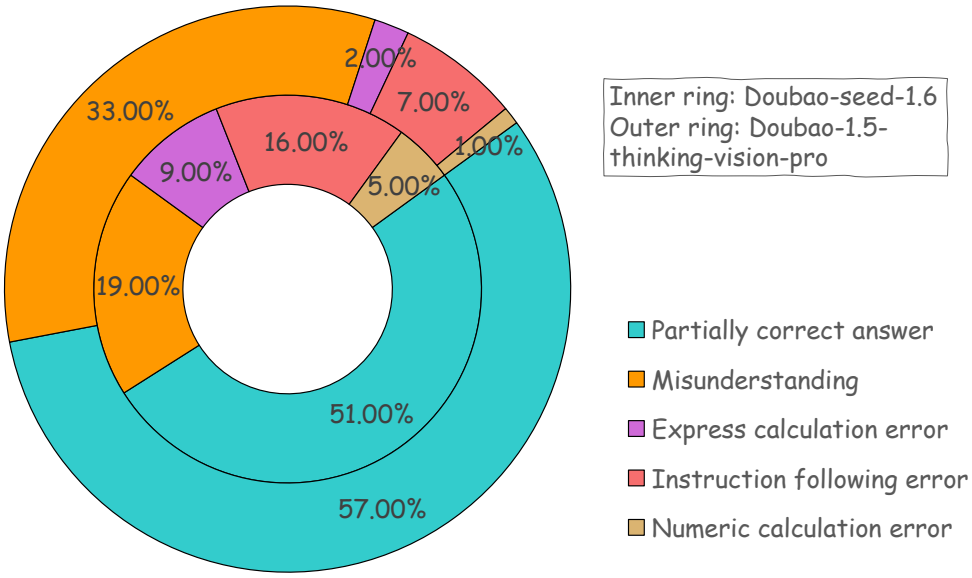


Figure 10: Error type distribution of Doubao.

H Illustrative Examples of Error Types

Example 1 (Express calculation error). Solve the differential equation:

$$\frac{dy}{dx} = 2 + \sin(y - 2x),$$

with initial condition $y(0) = -\frac{\pi}{2}$.

Reference Solution: Substitute $z = y - 2x \Rightarrow y = z + 2x$:

$$\frac{dy}{dx} = \frac{dz}{dx} + 2 = 2 + \sin(z) \quad \Rightarrow \quad \frac{dz}{dx} = \sin(z).$$

Integrate:

$$\int \frac{1}{\sin z}\, dz = \int dx \quad \Rightarrow \quad \ln\left|\tan\left(\frac{z}{2}\right)\right| = x + C.$$

Back-substitute $z = y - 2x$:

$$\ln\left|\tan\left(\frac{y - 2x}{2}\right)\right| = x + C.$$

Using $y(0) = \frac{\pi}{2}$:

$$\ln\left(\tan\frac{\pi/4}{1}\right) = C \quad \Rightarrow \quad C = 0.$$

Final solution:

$$\ln\left|\tan\left(\frac{y - 2x}{2}\right)\right| = x.$$

Student Solution: Let $v = y - 2x$, then:

$$\frac{dv}{dx} = \frac{dy}{dx} - 2.$$

Substitute into the original equation:

$$\frac{dv}{dx} + 2 = 2 + \sin(v) \quad \Rightarrow \quad \frac{dv}{dx} = \sin(v).$$

Separate variables and integrate:

$$\int \csc(v)\, dv = \int dx \quad \Rightarrow \quad \ln|\tan(v/2)| = x + C.$$

Exponentiating:

$$\tan\left(\frac{v}{2}\right) = Ke^x, \quad K = \pm e^C.$$

Substitute $v = y - 2x$:

$$\tan\left(\frac{y - 2x}{2}\right) = Ke^x.$$

Apply $y(0) = -\frac{\pi}{2}$:

$$\tan\left(-\frac{\pi}{4}\right) = K \quad \Rightarrow \quad K = -1.$$

Thus:

$$\tan\left(\frac{y - 2x}{2}\right) = -e^x.$$

Taking arctangent:

$$\frac{y - 2x}{2} = -\arctan(e^x) \quad \Rightarrow \quad y = 2x - 2\arctan(e^x).$$

Error Type: express calculation error.

Comment: While the method and integration were correct, the final expression does not match the reference form. The student's result introduces an explicit y in terms of $\arctan(e^x)$, whereas the reference keeps the solution in logarithmic–tangent form. This difference in algebraic representation is classified as an express calculation error.

Example 2 (Misunderstanding). Estimate the sum S of the series

$$S = \sum_{n=1}^{\infty} \frac{1}{n^2}$$

using the first 10 terms (i.e., with $n = 10$).

Reference Solution: Using the integral test remainder bounds:

$$\int_{n+1}^{\infty} \frac{1}{x^2}\, dx \le R_n \le \int_{n}^{\infty} \frac{1}{x^2}\, dx.$$

So for $n = 10$:

$$R_{10} \le \int_{10}^{\infty} \frac{1}{x^2}\, dx = \left[-\frac{1}{x}\right]_{10}^{\infty} = \frac{1}{10},$$

$$R_{10} \ge \int_{11}^{\infty} \frac{1}{x^2}\, dx = \frac{1}{11}.$$

Thus:

$$S_{10} + \frac{1}{11} \le S \le S_{10} + \frac{1}{10}.$$

From numeric calculation:

$$S_{10} \approx 1.54977 \quad \Rightarrow \quad 1.64068 \le S \le 1.64977.$$

A good estimate is the midpoint: $S \approx 1.6452$. Actually, from Fourier analysis:

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} \approx 1.64493.$$

Student Solution: To estimate the sum $S = \sum_{n=1}^{\infty} \frac{1}{n^2}$ using the first 10 terms, we compute the 10th partial sum $S_{10} = \sum_{n=1}^{10} \frac{1}{n^2}$.

Calculating each term and summing them:

- $n = 1$: $\frac{1}{1^2} = 1.0 \to$ cumulative sum: 1.0
- $n = 2$: $\frac{1}{2^2} = 0.25 \to$ cumulative sum: 1.25
- $n = 3$: $\frac{1}{3^2} \approx 0.1111111 \to$ cumulative sum: 1.3611111
- $n = 4$: $\frac{1}{4^2} = 0.0625 \to$ cumulative sum: 1.4236111
- $n = 5$: $\frac{1}{5^2} = 0.04 \to$ cumulative sum: 1.4636111
- $n = 6$: $\frac{1}{6^2} \approx 0.0277778 \to$ cumulative sum: 1.4913889
- $n = 7$: $\frac{1}{7^2} \approx 0.0204082 \to$ cumulative sum: 1.5117971
- $n = 8$: $\frac{1}{8^2} = 0.015625 \to$ cumulative sum: 1.5274221
- $n = 9$: $\frac{1}{9^2} \approx 0.0123457 \to$ cumulative sum: 1.5397678
- $n = 10$: $\frac{1}{10^2} = 0.01 \to$ cumulative sum: 1.5497678

Thus, the estimate of S using the first 10 terms is approximately 1.549768.

Error Type: misunderstanding.

Comment: The student computed the partial sum correctly, but misunderstood the task: they were expected to use the integral test remainder bounds to estimate S with an error bound, rather than simply summing the first 10 terms. This demonstrates a misunderstanding of the problem requirements.

Example 3: Partially correct answer. Suppose that $X_1, X_2, \ldots$ is a sequence of random variables such that $X_1 \sim \mathrm{Uniform}(0,1)$, and for $n = 1, 2, \ldots$, the conditional distribution of $X_{n+1} \mid X_1, \ldots, X_n$ is uniform on $[0, cX_n]$ for $c \in (\sqrt{3}, 2)$. Compute $\mathbb{E}[X_n^r]$ for $r > 0$.

Reference solution. We compute recursively:

$$\begin{aligned}\mathbb{E}[X_{n+1}^r] &= \mathbb{E}\big[\,\mathbb{E}[X_{n+1}^r \mid X_n]\,\big] \\ &= \mathbb{E}\left(\frac{1}{cX_n}\int_0^{cX_n} y^r\,dy\right) \\ &= \mathbb{E}\left(\frac{(cX_n)^{r+1}}{(r+1)\,cX_n}\right) = \frac{c^r}{r+1}\,\mathbb{E}[X_n^r].\end{aligned}$$

Recursively applying:

$$\mathbb{E}[X_n^r] = \left(\frac{c^r}{r+1}\right)^n \mathbb{E}[X_1^r] = \left(\frac{c^r}{r+1}\right)^n \cdot \frac{1}{r+1}.$$

Student solution. To compute $\mathbb{E}[X_n^r]$ for $r > 0$, we use the law of total expectation and induction.

Base case ($n = 1$). $X_1 \sim \mathrm{Uniform}(0,1)$, so its density is $f(x) = 1$ for $x \in (0,1)$. Then

$$\mathbb{E}[X_1^r] = \int_0^1 x^r\,dx = \frac{1}{r+1}.$$

Recurrence relation ($n \geq 2$). For $n \geq 2$, $X_n \mid X_{n-1} = x$ is uniform on $[0, cx]$. The conditional expectation is

$$\mathbb{E}[X_n^r \mid X_{n-1} = x] = \int_0^{cx} y^r \cdot \frac{1}{cx}\,dy = \frac{1}{cx} \cdot \frac{(cx)^{r+1}}{r+1} = \frac{c^r x^r}{r+1}.$$

Hence,

$$\mathbb{E}[X_n^r] = \mathbb{E}\left[\frac{c^r}{r+1} X_{n-1}^r\right] = \frac{c^r}{r+1}\,\mathbb{E}[X_{n-1}^r].$$

Inductive proof. Let $a_n = \mathbb{E}[X_n^r]$. Then

$$a_n = \frac{c^r}{r+1} a_{n-1}, \qquad a_1 = \frac{1}{r+1}.$$

Solving:

$$n = 1:\ a_1 = \tfrac{1}{r+1}, \quad n = 2:\ a_2 = \tfrac{c^r}{(r+1)^2}, \quad n = 3:\ a_3 = \tfrac{c^{2r}}{(r+1)^3}.$$

By induction, assume $a_k = \dfrac{c^{r(k-1)}}{(r+1)^k}$. Then

$$a_{k+1} = \frac{c^r}{r+1}\,a_k = \frac{c^{rk}}{(r+1)^{k+1}}.$$

Thus, for all $n \geq 1$,

$$\mathbb{E}[X_n^r] = \frac{c^{r(n-1)}}{(r+1)^n}.$$

Error type. Partially correct answer.

Comment. The student's recurrence setup is correct, but the final formula $\frac{c^{r(n-1)}}{(r+1)^n}$ does not match the reference $\left(\frac{c^r}{r+1}\right)^n \cdot \frac{1}{r+1}$, missing an extra factor $\frac{c^r}{r+1}$ for $n > 1$.

Example 4 (Numeric calculation error). Find the general solution of the following equation by the method of Variation of Parameters:

$$y'' - 7y' + 10y = 100x$$

Reference Solution: The homogeneous equation is:

$$y'' - 7y' + 10y = 0$$

Characteristic equation:

$$\lambda^2 - 7\lambda + 10 = (\lambda - 2)(\lambda - 5) \quad \Rightarrow \quad \lambda = 2, 5$$

Fundamental solutions:

$$y_1(x) = e^{2x}, \quad y_2(x) = e^{5x}$$

Compute Wronskian:

$$W[y_1, y_2](x) = e^{2x} \cdot 5e^{5x} - 2e^{2x} \cdot e^{5x} = 3e^{7x}$$

Let $g(x) = 100x$, apply variation of parameters:

$$\begin{aligned} y_p(x) &= -e^{5x} \int \frac{100se^{2s}}{3e^{7s}}\, ds + e^{2x} \int \frac{100se^{5s}}{3e^{7s}}\, ds \\ &= -\frac{100}{3} e^{5x} \int se^{-5s}\, ds + \frac{100}{3} e^{2x} \int se^{-2s}\, ds \\ &= -\frac{100}{3} e^{5x} \left(-\frac{1}{5} xe^{-5x} - \frac{1}{25} e^{-5x} \right) \\ &\quad + \frac{100}{3} e^{2x} \left(-\frac{1}{2} xe^{-2x} - \frac{1}{4} e^{-2x} \right) \\ &= \frac{20}{3} x + \frac{4}{3} - \frac{50}{3} x - \frac{25}{3} = -10x - 7 \end{aligned}$$

General solution:

$$y(x) = -10x - 7 + c_1 e^{2x} + c_2 e^{5x}$$

Student Solution: The student applied the Variation of Parameters method:

1. Complementary Solution: Found $y_c = C_1 e^{2x} + C_2 e^{5x}$ from the characteristic roots 2 and 5. 2. Wronskian: Correctly computed $W = 3e^{7x}$. 3. u_1' and u_2':

$$u_1' = -\frac{e^{5x} \cdot 100x}{3e^{7x}} = -\frac{100}{3} xe^{-2x}, \quad u_2' = \frac{e^{2x} \cdot 100x}{3e^{7x}} = \frac{100}{3} xe^{-5x}$$

4. Integration:

$$\begin{aligned} u_1 &= -\frac{100}{3} \int xe^{-2x} dx = \frac{50}{3} xe^{-2x} + \frac{25}{3} e^{-2x} \\ u_2 &= \frac{100}{3} \int xe^{-5x} dx = -\frac{20}{3} xe^{-5x} - \frac{4}{3} e^{-5x} \end{aligned}$$

5. Particular Solution:

$$\begin{aligned} y_p &= \left(\frac{50}{3} xe^{-2x} + \frac{25}{3} e^{-2x} \right) e^{2x} \\ &\quad + \left(-\frac{20}{3} xe^{-5x} - \frac{4}{3} e^{-5x} \right) e^{5x} \end{aligned}$$

Simplifying:

$$y_p = \frac{50}{3} x + \frac{25}{3} - \frac{20}{3} x - \frac{4}{3} = 10x + 7$$

6. General Solution:

$$y = C_1 e^{2x} + C_2 e^{5x} + 10x + 7$$

Error Type: numeric calculation error.

Comment: The student followed the correct Variation of Parameters procedure but made a sign error when combining the two terms in y_p. This caused the particular solution to be $10x+7$ instead of $-10x-7$, leading to a numeric calculation error in the final answer.

Example 5 (Instruction following error). Find all values of c so that $\lambda = 2$ is an eigenvalue of the matrix

$$A = \begin{bmatrix} 4 & -3 \\ 4 & c \end{bmatrix}.$$

Clearly circle your answer below:

(i) $c = -3$ only (ii) $c = -4$ only (iii) $c = 4$ only (iv) $c = -6$ only
(v) All c except -4 (vi) All c except -6 (vii) All c except 6

Reference Solution: We compute:

$$\det(A - 2I) = \det \begin{bmatrix} 2 & -3 \\ 4 & c-2 \end{bmatrix} = 2(c-2) + 12$$

Setting this equal to 0:

$$2(c-2) + 12 = 0 \quad \Rightarrow \quad 2c - 4 + 12 = 0$$

$$2c = -8 \quad \Rightarrow \quad c = -4$$

Correct choice: (ii).

Student Solution: To determine the values of c for which $\lambda = 2$ is an eigenvalue of A, we require:

$$\det(A - \lambda I) = 0.$$

First, compute:

$$A - 2I = \begin{bmatrix} 4-2 & -3 \\ 4 & c-2 \end{bmatrix} = \begin{bmatrix} 2 & -3 \\ 4 & c-2 \end{bmatrix}.$$

Determinant:

$$\det(A - 2I) = (2)(c-2) - (-3)(4) = 2(c-2) + 12$$

Simplifying:

$$\det(A - 2I) = 2c - 4 + 12 = 2c + 8$$

Setting equal to zero:

$$2c + 8 = 0 \quad \Rightarrow \quad c = -4.$$

Thus, the only value is $c = -4$.

Error Type: instruction following error.

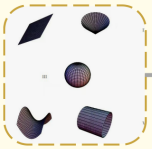
Comment: While the computation of c was correct, the final answer was not provided in the required multiple-choice format (i.e., option (ii) should have been explicitly circled as per instructions). This constitutes an instruction-following error.

I Case Study: Comparative Analysis of Answer Combinations across Three Image Versions in GPT–5 and Qwen3–VL–235B–A22B

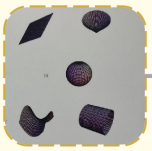
From top to bottom are the original image version, the printed version, and the hand-drawn version of GPT-5 or Qwen3-VL-235B-A22B.

Example of GPT-5

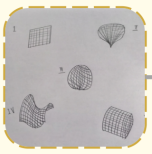
Question: Match the parameterized surface \\(\\vec{r}(u,v) = \\langle \\cos(u)\\sin(v), \\sin(u)\\sin(v), \\cos(v) \\rangle \\) with the corresponding picture and implicit surface equation.
Answer: C→III→x^2 + y^2 + z^2 = 1



\nCompute the Cartesian relation satisfied by the coordinates:\nx = cos(u) sin(v), y = sin(u) sin(v), z = cos(v).\n\nThen\nx^2 + y^2 + z^2\n= cos^2(u) sin^2(v) + sin^2(u) sin^2(v) + cos^2(v)\n= sin^2(v)(cos^2 u + sin^2 u) + cos^2(v)\n= sin^2(v) + cos^2(v)\n= 1.\n\nThus the surface is the unit sphere x^2 + y^2 + z^2 = 1. The parameterization with u ∈ [0, 2π), v ∈ [0, π] covers the entire sphere, matching the spherical picture labeled III.\n.
Final Answer: Picture III; implicit surface: x^2 + y^2 + z^2 = 1



\nCompute the squared norm of the position vector:\n\\[\n\\|\\vec r(u,v)\\|^2\n= (\\cos u \\sin v)^2 + (\\sin u \\sin v)^2 + (\\cos v)^2\n= \\sin^2 v(\\cos^2 u + \\sin^2 u) + \\cos^2 v\n= \\sin^2 v + \\cos^2 v\n= 1.\n\\]\nHence every point of the surface lies on the unit sphere centered at the origin, with implicit equation \\(x^2+y^2+z^2=1\\). The parameters \\(u\\) (azimuth) and \\(v\\) (polar angle) cover the sphere as in standard spherical coordinates (e.g., \\(u\\in[0,2\\pi]\\), \\(v\\in[0,\\pi]\\)). Therefore, the picture corresponding to this surface is the sphere (picture III).\n
Final Answer: \nPicture III; implicit equation \\(x^2 + y^2 + z^2 = 1\\

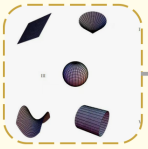


\nLet x = cos(u) sin(v), y = sin(u) sin(v), z = cos(v). Then\nx^2 + y^2 = sin^2(v)(cos^2(u) + sin^2(u)) = sin^2(v),\nand hence x^2 + y^2 + z^2 = sin^2(v) + cos^2(v) = 1.\nThus the surface is the unit sphere S^2. The usual parameter domain 0 ≤ u < 2π, 0 ≤ v ≤ π covers the entire sphere (with the usual degeneracy at the poles), producing the characteristic grid of longitudes and latitudes. Among the images, the spherical grid corresponds to picture III.\n
Final Answer: Picture III; implicit equation x^2 + y^2 + z^2 = 1

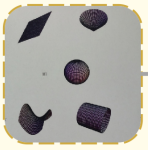
Figure 11: GPT-5 Original version is right, printint version is right and drawing version is right

Example of Qwen3-VL

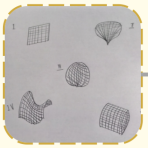
Question: Match the parameterized surface \\(\\vec{r}(u,v) = \\langle \\cos(u)\\sin(v), \\sin(u)\\sin(v), \\cos(v) \\rangle \\) with the corresponding picture and implicit surface equation.
Answer: C→III→x^2 + y^2 + z^2 = 1



\nCompute the Cartesian relation satisfied by the coordinates:\nx = cos(u) sin(v), y = sin(u) sin(v), z = cos(v).\n\nThen\nx^2 + y^2 + z^2\n= cos^2(u) sin^2(v) + sin^2(u) sin^2(v) + cos^2(v)\n= sin^2(v)(cos^2 u + sin^2 u) + cos^2(v)\n= sin^2(v) + cos^2(v)\n= 1.\n\nThus the surface is the unit sphere x^2 + y^2 + z^2 = 1. The parameterization with u ∈ [0, 2π), v ∈ [0, π] covers the entire sphere, matching the spherical picture labeled III.\n
Final Answer: Picture III; implicit surface: x^2 + y^2 + z^2 = 1.



\nCompute the squared norm of the position vector:\n\\[\n\\|\\vec{r}(u,v)\\|^2\n= (\\cos u \\sin v)^2 + (\\sin u \\sin v)^2 + (\\cos v)^2\n= \\sin^2 v(\\cos^2 u + \\sin^2 u) + \\cos^2 v\n= \\sin^2 v + \\cos^2 v\n= 1.\n\\]\nHence every point of the surface lies on the unit sphere centered at the origin, with implicit equation \\(x^2+y^2+z^2=1\\). The parameters \\(u\\) (azimuth) and \\(v\\) (polar angle) cover the sphere as in standard spherical coordinates (e.g., \\(u\\in[0,2\\pi]\\), \\(v\\in[0,\\pi]\\)). Therefore, the picture corresponding to this surface is the sphere (picture III).\n
Final Answer: ∠CAB = 26°

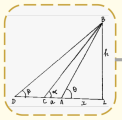


\nThe given parameterization is:\n\n\\[\n\\vec{r}(u,v) = \\langle \\cos(u)\\sin(v), \\sin(u)\\sin(v), \\cos(v) \\rangle\n\\]\n\nThis is a standard parameterization of the **unit sphere** in spherical coordinates, where:\n\n- \\(u \\in [0, 2\\pi) \\) corresponds to the azimuthal angle (longitude),\n- \\(v \\in [0, \\pi] \\) corresponds to the polar angle (colatitude, measured from the positive z-axis).\n\nLet's denote the components as:\n\n\\[\nx = \\cos(u)\\sin(v), \\quad y = \\sin(u)\\sin(v), \\quad z = \\cos(v)\n\\]\n\nWe now derive the implicit equation by computing \\(x^2 + y^2 + z^2 \\):\n\n\\[\nx^2 + y^2 = \\cos^2(u)\\sin^2(v) + \\sin^2(u)\\sin^2(v) = \\sin^2(v)(\\cos^2(u) + \\sin^2(u)) = \\sin^2(v)\n\\]\n\nThen,\n\n\\[\nx^2 + y^2 + z^2 = \\sin^2(v) + \\cos^2(v) = 1\n\\]\n\nThus, the implicit equation is:\n\n\\[\nx^2 + y^2 + z^2 = 1\n\\]\n\nThis is the equation of the **unit sphere** centered at the origin.\n\nNow, examining the provided images:\n\n- Image I: A flat grid — likely a plane or rectangle.\n- Image II: A cone-like or funnel-shaped surface — not a sphere.\n- Image III: A spherical surface with grid lines — matches the parameterization.\n- Image IV: A saddle or warped surface — likely a hyperbolic paraboloid.\n- Image V: A cylinder — clearly not a sphere.\n\nTherefore, the correct match is **Image III**.\n
Final Answer: III, \\(x^2 + y^2 + z^2 = 1 \\)

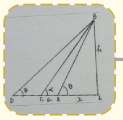
Figure 12: Qwen3-VL-A235B-a22B original version is right, printing version is right and drawing version is right

Example of GPT-5

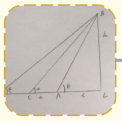
Question: Two stations due south of a tower, which leans towards north are at distances a and b from its foot. If $\\alpha$ and $\\beta$ be the elevations of the top of the tower from the situation, prove that its inclination $\\theta$ to the horizontal is given by\n\\[\\cot \\theta = \\frac{b \\cot \\alpha - a \\cot \\beta}{b - a} \\]
Answer: \\boxed{\\frac{b \\cot \\alpha - a \\cot \\beta}{b - a}}



\nLet O be the foot of the tower and T its top. Since the tower leans towards north, the horizontal projection of T from O is towards the north. Denote by:\n- h = vertical height of T above O,\n- x = horizontal displacement of T from O towards north,\n- θ = inclination of OT to the horizontal. Then tan θ = h/x and hence cot θ = x/h.\n\nTake two observation points A and D due south of O at distances a and b, respectively; i.e., A and D are on the same horizontal line south of O, at coordinates (−a, 0) and (−b, 0) if O is at (0, 0), while T is at (x, h). The horizontal distance from A to the vertical through T is a + x; from D it is b + x. Therefore the angles of elevation satisfy\ntan α = h/(a + x), tan β = h/(b + x).\n\nHence\ncot α = (a + x)/h, cot β = (b + x)/h.\n\nCompute\nb cot α − a cot β\n= b(a + x)/h − a(b + x)/h\n= (ab + bx − ab − ax)/h\n= (b − a) x/h.\n\nDividing by b − a gives\n(b cot α − a cot β)/(b − a) = x/h = cot θ.\n\nThis completes the proof.\n
Final Answer: \nThis completes the proof.\n



\nLet A be the foot of the tower and B its top. Since the tower leans towards the north, let the horizontal displacement of the top from the foot be x towards the north, and let the vertical height of the top above the ground be h. Then the inclination $\\theta$ of the tower to the horizontal satisfies\n\\[\n\\tan\\theta=\\frac{h}{x}\\quad\\Rightarrow\\quad \\cot\\theta=\\frac{x}{h}.\n\\]\n\nLet C and D be the two stations due south of A with $AC=a$ and $AD=b$ (so $b>a$). From C and D the horizontal distances to the vertical projection of B are $x+a$ and $x+b$, respectively. Hence the angles of elevation satisfy\n\\[\n\\tan\\alpha=\\frac{h}{x+a},\\qquad \\tan\\beta=\\frac{h}{x+b}.\n\\]\nTaking cotangents gives\n\\[\n\\cot\\alpha=\\frac{x+a}{h}=\\frac{x}{h}+\\frac{a}{h},\\qquad \n\\cot\\beta=\\frac{x+b}{h}=\\frac{x}{h}+\\frac{b}{h}.\n\\]\nMultiply the first by b and the second by a, then subtract:\n\\[\nb\\cot\\alpha - a\\cot\\beta\n= b\\left(\\frac{x}{h}+\\frac{a}{h}\\right)-a\\left(\\frac{x}{h}+\\frac{b}{h}\\right)\n= (b-a)\\frac{x}{h}.\n\\]\nTherefore,\n\\[\n\\frac{x}{h}=\\frac{b\\cot\\alpha - a\\cot\\beta}{b-a}.\n\\]\nSince $\\frac{x}{h}=\\cot\\theta$, we obtain\n\\[\n\\cot \\theta = \\frac{b \\cot \\alpha - a \\cot \\beta}{b - a},\n\\]\nas required
Final Answer: ∠CAB = 26°



\nLet O be the foot and T the top of the tower. Since the tower leans towards the north, in the vertical north–south plane we may write the coordinates:\n- O at x = 0,\n- T at x = d (north of O), height h.\n\nThus the tower OT makes an angle θ with the horizontal and\ncot θ = d/h.\n\nPlace two stations S_a and S_b due south of O at distances a and b respectively, i.e., at x = −a and x = −b.\n\nFrom S_a the angle of elevation of T is α, hence\ntan α = h/(a + d)\n⇔ h cot α = a + d. (1)\n\nFrom S_b the angle of elevation is β, hence\ntan β = h/(b + d)\n⇔ h cot β = b + d. (2)\n\nFrom (1) and (2),\nb(h cot α − a) − a(h cot β − b) = b d − a d = (b − a) d,\nso\n(b − a) d = h(b cot α − a cot β).\n\nDivide by h(b − a) to obtain\nd/h = (b cot α − a cot β)/(b − a).\n\nBut d/h = cot θ, since θ is the inclination of the tower to the horizontal. Therefore,\ncot θ = (b cot α − a cot β)/(b − a),\nas was to be proved.\n
Final Answer: \ncot θ = (b cot α − a cot β)/(b − a)

Figure 13: GPT-5 original version is right, printing version is right but drawing version is wrong

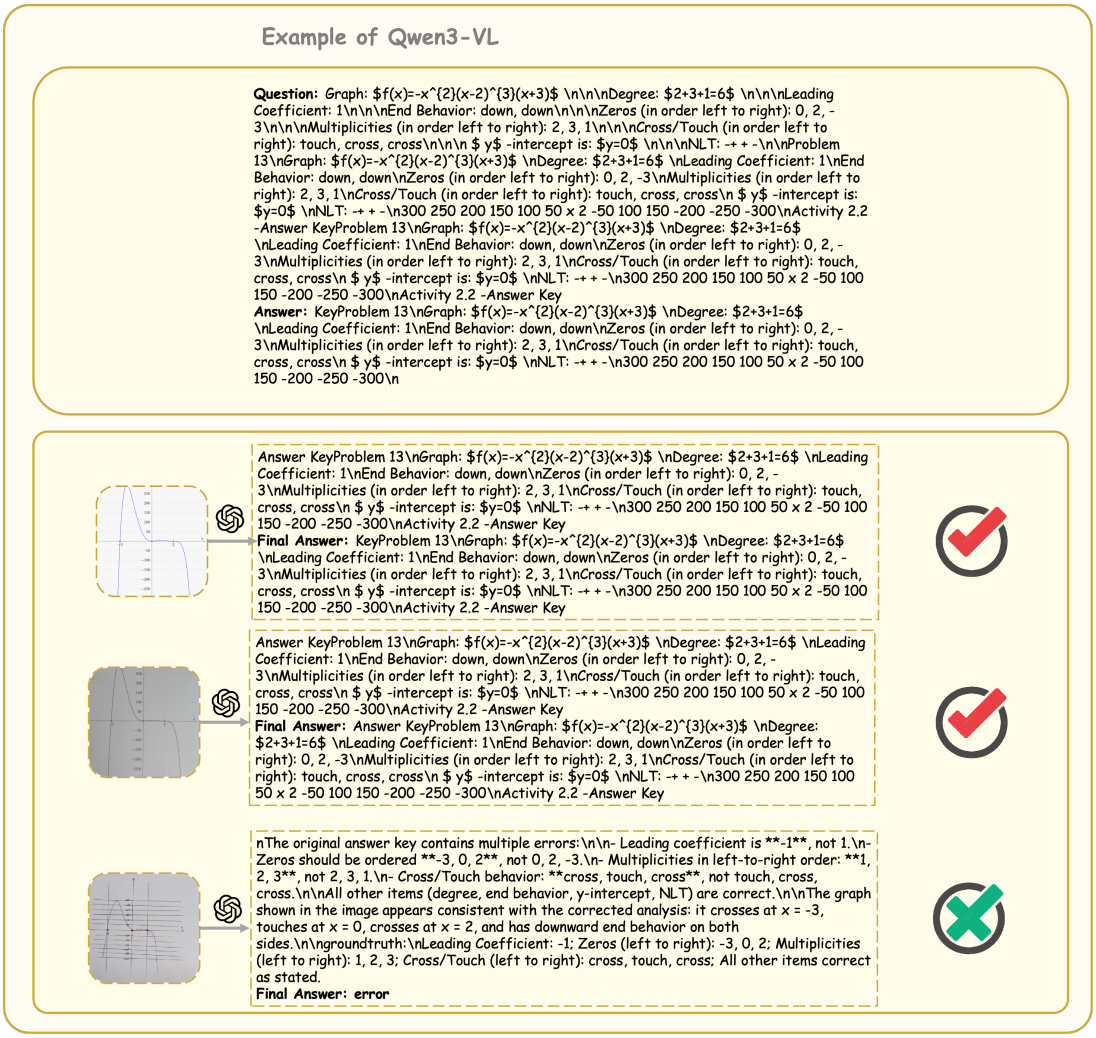
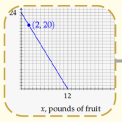


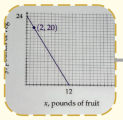
Figure 14: Qwen3-VL-A235B-A22B original version is right, printing version is right but drawing version is wrong

Example of QwenVL-3

Question: We have $24 to spend on vegetables and fruit. A pound of vegetables costs $1 and a pound of fruit costs $2. The number of pounds of vegetables we can afford, y, is a function of the number of pounds of fruit we buy, x. On the axes below:\n• Graph your equation.\n• Label each axis by writing the name of the variable and its units along the axis.\n• Label the coordinates at the vertical and horizontal intercepts.\n• Draw a dot at the point on your graph corresponding to a purchase of 2 pounds of fruit and label its coordinates
Answer: Graph of y = -2x + 24, with x-axis labeled as 'pounds of fruit', y-axis labeled as 'pounds of vegetables', intercepts (0, 24) and (12, 0) marked, point (2, 20) labeled



Let x be pounds of fruit and y be pounds of vegetables. With a total budget of $24, the cost equation is\n2x + 1y = 24.\nSolving for y gives the linear function:\ny = 24 – 2x.\n\nKey features for the graph:\n- Axes labels:\n - Horizontal axis: x, pounds of fruit.\n - Vertical axis: y, pounds of vegetables.\n- Intercepts:\n - Vertical intercept (x = 0): y = 24 → point (0, 24).\n - Horizontal intercept (y = 0): 24 – 2x = 0 → x = 12 → point (12, 0).\n- Plot the straight line segment between (0, 24) and (12, 0). The slope is –2.\n- Point for 2 pounds of fruit:\n - Substitute x = 2 into y = 24 – 2x: y = 24 – 4 = 20 → point (2, 20).\n - Place a dot at (2, 20) and label it.\n\nDomain/feasible region for purchases is x ≥ 0, y ≥ 0, so the relevant graph is the segment of the line between (0, 24) and (12, 0).\n
Final Answer: Equation: y = 24 – 2x\nAxis labels: x = pounds of fruit, y = pounds of vegetables\nIntercepts: (0, 24) and (12, 0)\nPoint for 2 lb fruit: (2, 20)



Let x be pounds of fruit and y be pounds of vegetables.\nCost constraint: 2x + 1·y = 24, so y = 24 – 2x.\nDomain (nonnegative quantities): 0 ≤ x ≤ 12 and 0 ≤ y ≤ 24.\n\nGraph:\n- Plot the line y = 24 – 2x on the first quadrant only.\n- Label axes:\n - Horizontal axis: x, pounds of fruit.\n - Vertical axis: y, pounds of veg.\n\nIntercepts:\n- Vertical intercept (x = 0): y = 24 → point (0, 24).\n- Horizontal intercept (y = 0): 2x = 24 → x = 12 → point (12, 0).\nLabel both coordinates on the graph.\n\nSpecific purchase:\n- For x = 2 pounds of fruit, y = 24 – 2(2) = 20.\n- Place and label a point at (2, 20).\n\nThe graph is a straight line with slope –2 descending from (0, 24) to (12, 0), with the marked point (2, 20).\n
Final Answer: y = 24 – 2x, 0 ≤ x ≤ 12.\nIntercepts: (0, 24) and (12, 0).\nMarked point: (2, 20).\nAxis labels: x (pounds of fruit), y (pounds of veg)

n- Let x be the number of pounds of fruit and y be the number of pounds of vegetables.\n- The total cost must equal 24 dollars: 2x + 1y = 24.\n- Solve for y to express it as a function of x: y = 24 – 2x.\n- Domain and range from nonnegativity: x ∈ [0,12] and y ∈ [0,24].\n\nHow to graph:\n- Label the horizontal axis "x, pounds of fruit" and the vertical axis "y, pounds of vegetables (lbs)".\n- Plot the vertical intercept at x = 0: y = 24 → (0, 24). Label it "(0, 24)".\n- Plot the horizontal intercept where y = 0: 24 – 2x = 0 → x = 12 → (12, 0). Label it "(12, 0)".\n- Draw the straight line segment between (0,24) and (12,0). This is the graph of y = 24 – 2x.\n- For a purchase of 2 pounds of fruit (x = 2), y = 24 – 2·2 = 20. Plot a dot at (2, 20) and label it "(2, 20)".\n\nThis line has slope –2 and shows all affordable combinations of fruit and vegetables under the $24 budget.\n
Final Answer:n- Let x be the number of pounds of fruit and y be the number of pounds of vegetables.\n- The total cost must equal 24 dollars: 2x + 1y = 24.\n- Solve for y to express it as a function of x: y = 24 – 2x.\n- Domain and range from nonnegativity: x ∈ [0,12] and y ∈ [0,24].\n\nHow to graph:\n- Label the horizontal axis "x, pounds of fruit" and the vertical axis "y, pounds of vegetables (lbs)".\n- Plot the vertical intercept at x = 0: y = 24 → (0, 24). Label it "(0, 24)".\n- Plot the horizontal intercept where y = 0: 24 – 2x = 0 → x = 12 → (12, 0). Label it "(12, 0)".\n- Draw the straight line segment between (0,24) and (12,0). This is the graph of y = 24 – 2x.\n- For a purchase of 2 pounds of fruit (x = 2), y = 24 – 2·2 = 20. Plot a dot at (2, 20) and label it "(2, 20)".\n\nThis line has slope –2 and shows all affordable combinations of fruit and vegetables under the $24 budget.\n

Figure 15: Qwen3-VL-A235B-A22B original version is wrong, printing version is wrong and drawing version is wrong