

# VisAidMath: Benchmarking Visual-Aided Mathematical Reasoning

Jingkun Ma<sup>1</sup> Runzhe Zhan<sup>1</sup> Yang Li<sup>1</sup>

Di Sun<sup>2</sup> Hou Pong Chan<sup>3</sup> Lidia S. Chao<sup>1</sup> Derek F. Wong<sup>1\*</sup>

<sup>1</sup>NLP<sup>2</sup>CT Lab, Department of Computer and Information Science, University of Macau

<sup>2</sup> Department of Mathematics, University of Macau

<sup>3</sup> DAMO Academy, Alibaba Group

nlp2ct.{jingkun, runzhe, liyang}@gmail.com,

mc25219@um.edu.mo, houpong.chan@alibaba-inc.com

lidiasc@um.edu.mo, derekfw@um.edu.mo

## Abstract

A hallmark of advanced artificial intelligence is the capacity to progress from passive visual perception to the strategic modification of visual information to facilitate complex reasoning. This advanced capability, however, remains critically underdeveloped in current Large Multi-modal Models (LMMs). The deficiency is often masked by evaluation metrics that prioritize final-answer accuracy, creating an illusion of competence where genuine reasoning is absent. Using the domain of geometric problem-solving as a precise instrument, we probe this issue through tasks that require constructing visual aids. To this end, we introduce **VisAidMath**, a challenging benchmark, and our novel Three-Layered Funnel Evaluation Framework. This framework moves beyond simple accuracy (ACCU) to scrutinize the generation of valid visual aids (PVA) and the soundness of subsequent reasoning steps (SPRS). Our extensive experiments on state-of-the-art models, including Doubao-Seed-1.6 and o4, reveal a profound “Reasoning Illusion”. We observe that high surface-level accuracy conceals a catastrophic failure in the models’ ability to produce valid visual aids or to reason from them. Our findings expose a fundamental schism between visual perception and logical deduction in modern LMMs. We provide a public evaluation platform on CodaBench<sup>1</sup> and release the project homepage<sup>2</sup>.

## 1 Introduction

Mathematical problem-solving (MPS) remains a significant hurdle for Large Language Models (LLMs) (Shao et al., 2024; Guan et al., 2025), and Large Multi-modal Models (LMMs) (Hendrycks et al., 2021; Lewkowycz et al., 2022; Wu et al., 2020, 2021; Shi et al., 2024; Peng et al., 2024).

\* Corresponding author.

<sup>1</sup><https://www.codabench.org/competitions/7634/>

<sup>2</sup><https://nlp2ct.github.io/VisAidMathHomepage/>

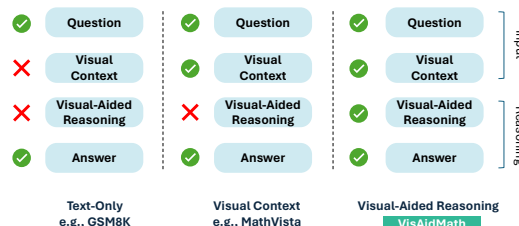


Figure 1: Comparison between VisAidMath and other benchmarks. Our work particularly focuses on utilization of explicit and implicit visual context during reasoning process.

The complexity intensifies when problems are presented in a multi-modal format (Lindström and Abraham, 2022; Masry et al., 2022). While some research has explored multi-modal MPS by incorporating visual contexts (Masry et al., 2022; Lu et al., 2023), these efforts predominantly focus on traditional vision-language task paradigms. Consequently, they tend to evaluate text-only reasoning steps, often underutilizing the visual information crucial for complex reasoning (Su et al., 2025). This creates a significant gap between evaluating surface-level “Think about Images” and “Think with Images” (Su et al., 2025), a gap our work aims to address as illustrated in Figure 1.

To bridge this gap, we introduce **VisAidMath**, a benchmark specifically designed to compel and evaluate the process of visual-aided mathematical reasoning. Drawing from diverse, high-quality sources, VisAidMath consists of 1,200 problems structured to necessitate generation or utilization of visual aids. This design moves beyond simple visual comprehension to directly target a model’s ability to perform complex spatial and logical deductions grounded in visual context. Crucially, VisAidMath provides a unified testbed to evaluate model capabilities across the full evolutionary spectrum of “Thinking with Images” paradigm (Su et al., 2025). Complex capabilities are better understood through structured, multi-level progression

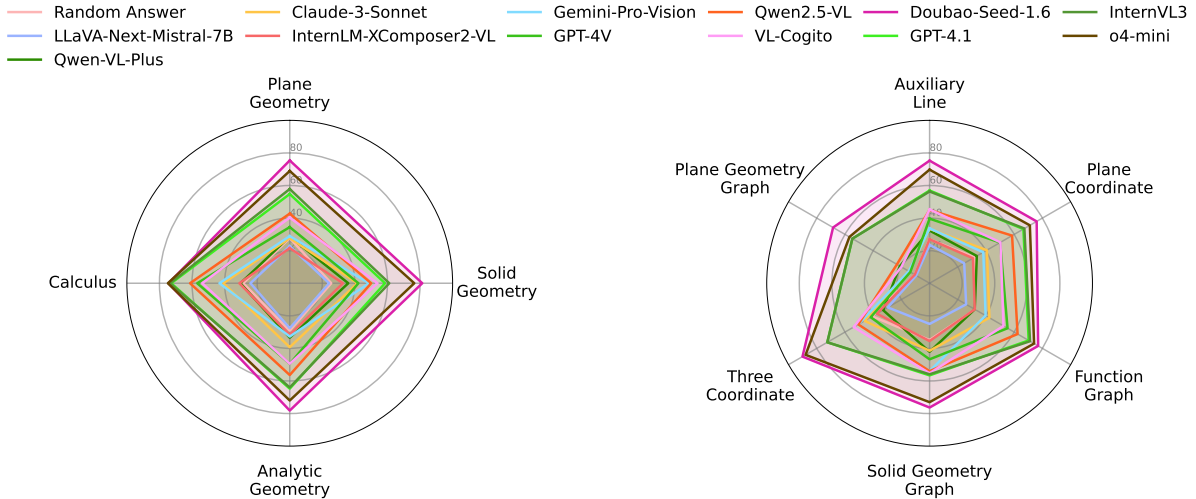


Figure 2: Accuracies of all LMM on visual-aided mathematical reasoning task across four branches and six visual aids.

rather than flat evaluation alone (Chen et al., 2023). Problems within our benchmark can be explored through the selection of analytical tools (Hu et al., 2022; Wu et al., 2024; Liu et al., 2025; Qi et al., 2024), the programmatic creation of visual aids like auxiliary lines (Gupta and Kembhavi, 2023; Fu et al., 2025b; Chervonyi et al., 2025; Weng et al., 2025), or by engaging in intrinsic spatial imagination to foresee the solution path (Team, 2024; Guo et al., 2025b; Zhao et al., 2025). Such diversity also suggests that model competence should be examined under structured and potentially dynamic difficulty regimes (Chen et al., 2024).

To demonstrate benchmark’s unique value, we propose a three-layered funnel evaluation framework that moves beyond standard accuracy (ACCU), which often masks procedural flaws. This framework assesses Process-Verified Accuracy (PVA), filtering out answers from flawed reasoning, and Solution Process Robustness Score (SPRS) to quantify the fine-grained quality of solution. Applying this rigorous evaluation reveals a universal “reasoning illusion”: a dramatic collapse from high accuracy to poor process quality across all tested models. Inspired by visual illusion (Zhang et al., 2025), we use **reasoning illusion** to describe a failure mode in LMMs where seemingly coherent reasoning is in fact disconnected from the problem’s actual constraints, especially its visual constraints. This illusion is most pronounced on the core tasks of VisAidMath, where even top models like Doubao-Seed-1.6 and o4-mini show a staggering performance drop. This collapse reveals a fundamental weakness in generating and utiliz-

ing visual aids, a deficiency completely missed by standard metrics.

This paper’s contributions are threefold: (1) We introduce **VisAidMath**, a novel benchmark that mandates visual-aided reasoning for mathematical problem-solving. (2) We propose a new **evaluation framework** that uncovers the widespread “reasoning illusion” in current SOTA models. (3) Through comprehensive analysis, we provide quantitative proof that VisAidMath is uniquely effective at exposing these deep-seated reasoning failures, thereby establishing its value and pinpointing critical areas for the future development of more robust and reliable multi-modal models.

## 2 VisAidMath

### 2.1 Data Creation

**Principles** A typical problem within our *VisAidMath* benchmark comprises four parts: **Visual Context (C)**, **Question (Q)**, **Visual Aids (V)**, and **Answer (A)**. The main task involves prompting the model to generate visual aids that assist in mathematical reasoning, a key distinction from other benchmarks as detailed in appendix K. While the visual context may be optional, spatial descriptions are included as essential data elements within the question. Given that many text-based LLMs lack image understanding or generation capabilities, we have additionally annotated precise captions for both the visual context and the visual aids through annotation. This allows us to extend the evaluation scenarios to models that are constructed with limited modality.

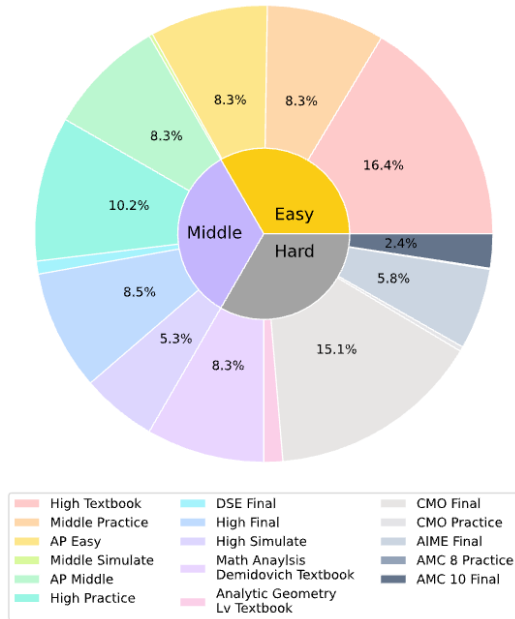


Figure 3: Distribution of data sources and difficulty levels.

**Data Sources and Categories** We collected the VisAidMath benchmark from both English and Chinese sources. All data sources were categorized to ensure a balanced range of difficulty. To ensure a range of difficulty levels, we categorized the data samples based on their sources into the following categories: Easy, Medium and Hard. We analyze the problem complexity of 16 data source following three difficulty levels: 1) Easy: Chinese High school entrance examination 2) Medium: Chinese College Entrance Examination 3) High: Chinese Mathematical Olympiad. Data distribution of each data source is shown in Figure 3. A comprehensive description of our data sources, collection methodology, metadata, and translation process is detailed in Appendix C.1 and C.2. To ensure diversity and balance, we manually collected and annotated a range of categories within the benchmark. Detailed categories and examples from different categorizations can be found in section D.

**Construction Pipeline** As depicted in Figure 7, we propose a construction pipeline for the VisAidMath dataset, which incorporates multi-round verification and dynamic quality control based on feedback. The dataset creation pipeline involves four key roles (see Appendix E.3). To prepare the benchmark for the wide English research community, we perform several steps. This includes machine translation, for which we selected DeepL and Baidu Translate after a manual sampling process

confirmed their high quality for technical content. Importantly, all translated samples are further reviewed and refined by human annotators to improve fluency and preserve mathematical meaning. We additionally conduct a post-refinement human evaluation on the final English data to verify translation quality. Details are provided in Appendix B.1. This is followed by data processing and release preparation. The detailed processes can be found in Appendix E, with the annotation process specified in Appendix E.1.

## 2.2 Benchmark Attributes

The distribution of data sources is presented in Figure 3, providing a comprehensive overview of the dataset’s origins. Additionally, the mathematical branches within the dataset exhibit a well-balanced distribution. This balance enables a broader exploration of diverse mathematical knowledge. Further details on other attributes can be found in Appendix C.

## 2.3 Task Definition

Our work introduces a series of novel mathematical reasoning tasks centered on the use of visual aids.

To formalize these tasks, we first define their basic components: the visual context (C), the textual question (Q), and the final answer (A). Crucially, a visual aid (V) in our framework is a **textual description** of a geometric construction.

We establish a baseline task, **General Reasoning (GR)**, and introduce two novel formulations based on how visual aids are utilized: **Direct Visual-Aided Reasoning (D-VAR)** tasks the model with generating a visual aid  $V_g$  to solve a problem. In contrast, **Indirect Visual-Aided Reasoning (I-VAR)** requires the model to leverage a provided one  $V_p$ . The formal definitions and core distinctions for these tasks are detailed in Table 1.

To accommodate language models that only accept textual inputs, the visual context ‘C’ is replaced by its textual description, denoted as  $C_{\text{txt}}$ . This creates text-only variants for each task, such as  $(C_{\text{txt}}, Q) \rightarrow A$  for the GR task. This approach ensures a fair comparison across both multimodal and text-only models, as the visual aids themselves are consistently represented as text in all settings.

As defined in Table 1, the introduction of the D-VAR and I-VAR tasks distinguishes VisAidMath from traditional benchmarks. Unlike tasks that only require understanding existing visual elements, our approach compels models to engage in a deeper,

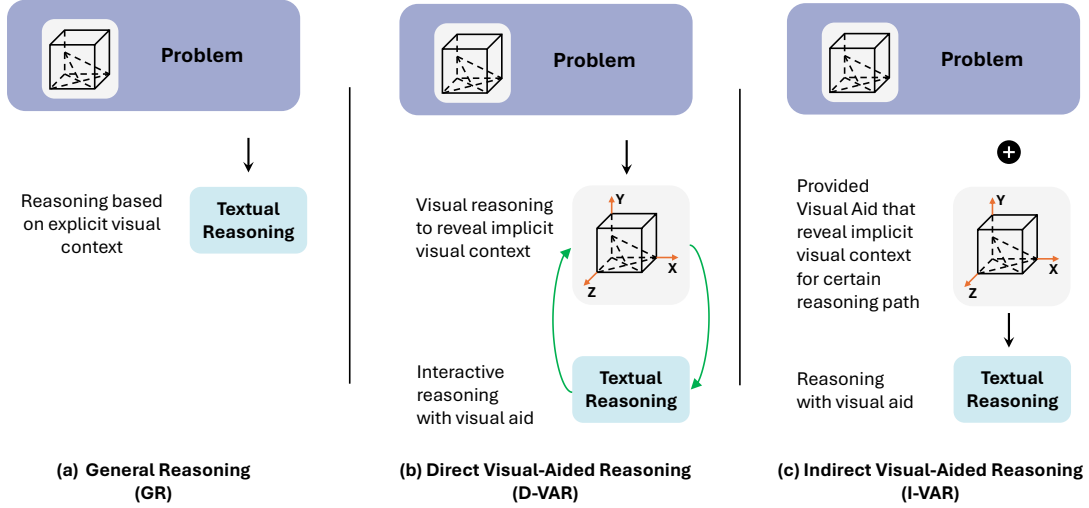


Figure 4: Comparison of different tasks: a) General Reasoning: provide MPS reasoning steps directly. b) Direct Visual-Aided Reasoning: create visual aids that disclose implicit visual context within problem, incorporating with textual reasoning to solve mathematical problem. c) Indirect Visual-Aided Reasoning: solve the mathematical problem based on given visual aids. Direct Visual-Aided reasoning require the model to perform visual reasoning for visual aids generation.

Table 1: Formal definitions of the reasoning tasks.  $V_g$  denotes a generated visual aid (output), while  $V_p$  denotes a provided visual aid (input).

Task Name	Abbr.	Formal Expression	Core Characteristic
General Reasoning	GR	$(C, Q) \rightarrow A$	Directly solves the problem.
Direct Visual-Aided Reasoning	D-VAR	$(C, Q) \rightarrow (V_g, A)$	<b>Generates</b> a visual aid to assist reasoning.
Indirect Visual-Aided Reasoning	I-VAR	$(C, Q, V_p) \rightarrow A$	<b>Utilizes</b> a provided visual aid for reasoning.

two-step reasoning process: first, planning or interpreting geometric constructions (the visual aids), and then executing the subsequent reasoning steps. This novel modality, focused on the generation and utilization of textual visual aids, allows us to specifically probe the spatial and logical planning capabilities of LMMs. We illustrate the distinctions between these tasks in Figure 4.

### 3 Experiments

#### 3.1 Models

To comprehensively evaluate the challenges posed by VisAidMath, we selected a wide spectrum of models, with a particular focus on large multi-modal models designed for complex reasoning. Our selection encompasses both leading open-source and proprietary systems. Our evaluation includes: 1) Open-source LLMs: Llama-2-7B (Touvron et al., 2023), Mistral-7B-Instruct-v0.2 (Jung et al., 2010); 2) Closed-source LLMs: GPT-3.5-turbo (OpenAI, 2023a),

GPT-4-turbo (OpenAI, 2023b); 3) Open-source LMMs: LLaVA-Next-Mistral-7B (Liu et al., 2024a), InternLM-XComposer2-VL (Dong et al., 2024), VL-Cogito (Yuan et al., 2025), Qwen2.5-VL-72B (Bai et al., 2025), InternVL3.5-38B (Wang et al., 2025); 4) Closed-source LMMs: Qwen-VL-Plus (Bai et al., 2023), Gemini-Pro-Vision (Google, 2023), Claude-3-Sonnet (Anthropic, 2024), GPT-4-Vision (OpenAI, 2023d), GPT-4.1 (OpenAI, 2023c), o4-mini (OpenAI, 2023e), and Doubao-Seed-1.6 (Guo et al., 2025a). This extensive selection allows for a robust analysis of performance trends across different model architectures, scales, and training paradigms. Detailed experimental settings and hyperparameters are provided in Appendix F.1.

#### 3.2 Three-Layered Funnel Evaluation

To move beyond surface-level correctness and enable a deeper assessment of mathematical reasoning, we introduce **The Three-Layered Funnel Evaluation Framework**. This hierarchical

Model	ALL	PLG	SDG	AYG	CAL	AXL	RTC	THC	PLG	SDG	FUG
<i>Heuristics Baselines</i>											
Random Answer	24.42	21.54	34.31	21.45	20.07	24.44	20.87	35.16	10.53	32.89	21.50
Frequent Answer	40.83	28.92	50.65	40.36	44.22	32.79	47.25	74.73	20.00	47.73	44.53
<i>Large Language Models (LLMs): Text-Only Input</i>											
Llama2-7B	26.83	21.85	34.64	30.55	20.75	26.68	25.23	39.56	11.58	30.26	26.49
Mistral-7b-Instruct-v0.2	27.42	27.38	30.72	27.64	23.81	27.57	28.21	28.57	11.58	27.63	26.87
GPT3.5	37.58	32.31	42.16	37.45	38.78	37.56	38.30	40.66	13.68	42.11	38.20
GPT4	51.92	41.54	52.29	50.91	63.95	45.75	54.59	60.44	23.16	53.29	61.23
<i>Large Multimodal Models (LMMs): Text-Only Input</i>											
LLaVA-Next-Mistral-7B	23.08	21.23	22.55	25.45	23.47	22.21	23.62	25.27	8.42	26.32	25.34
InternLM-XComposer2-VL	33.17	24.62	44.12	32.36	31.97	30.40	33.03	46.15	10.53	41.45	34.17
Qwen-VL-Plus	34.75	30.15	43.46	33.82	31.63	34.43	34.63	48.35	21.05	44.74	32.63
Gemini-Pro-Vision	38.42	31.08	48.37	31.27	42.86	34.72	37.84	49.45	18.95	51.97	39.54
Claude-3-Sonnet	38.58	31.38	43.46	39.27	40.82	36.66	40.14	46.15	14.74	43.42	42.23
GPT4V	47.00	35.08	47.06	50.55	56.80	41.43	50.69	48.35	15.79	47.37	55.66
<i>Large Multimodal Models (LMMs): Multimodal Input</i>											
LLaVA-Next-Mistral-7B	24.58	22.77	24.18	27.64	24.15	23.55	24.54	29.67	9.47	25.00	25.91
InternLM-XComposer2-VL	29.00	21.54	32.68	31.64	30.95	26.97	30.73	37.36	10.53	35.53	32.05
Qwen-VL-Plus	32.00	28.62	35.95	33.45	30.27	32.34	33.49	32.97	21.05	42.11	32.05
Gemini-Pro-Vision	38.33	28.92	48.69	32.73	43.20	33.68	38.07	50.55	14.74	53.95	39.73
Claude-3-Sonnet	37.08	27.69	41.50	39.27	40.82	33.38	40.60	46.15	14.74	41.45	42.42
GPT4V	45.33	34.46	42.16	49.45	56.80	39.64	50.00	41.76	13.68	46.71	55.28
VL-Cogito	49.17	40.31	53.92	53.74	49.45	45.31	53.85	52.40	55.26	50.23	20.00
Qwen2.5-VL-72B	52.25	42.77	50.00	61.22	56.36	45.01	50.55	62.38	53.95	58.49	23.16
GPT4.1	62.42	54.77	58.50	72.79	64.73	56.93	72.53	70.25	56.58	66.51	54.74
InternVL3.5-38B	63.92	57.85	61.11	73.47	64.00	56.33	72.53	71.21	55.92	67.20	54.74
o4-mini	73.00	68.92	76.47	74.83	72.00	69.75	87.91	74.09	73.03	71.10	56.84
Doubao-seed-1.6	77.33	75.38	81.37	74.49	78.18	75.26	90.11	76.97	76.32	75.92	68.42

Table 2: Accuracy scores on **Direct Visual-Aided Reasoning (D-VAR)** task upon **VisAidMath**. Meanings of all abbreviations are: ALL → overall accuracy, PLG → plane geometry, SDG → solid geometry, AYG → analytic geometry, CAL: calculus and functions. Visual Aids Types: AXL → auxiliary line, RTC → rectangular coordinate, THC → rectangular three-dimensional coordinate, PLG → plane geometry graph, SDG → solid geometry graph, FUG → function graph. The highest scores in chunk and in general have been emphasized with purple and green to facilitate comparison respectively.

methodology is designed to scrutinize model outputs with increasing depth and rigor, moving from surface-level answer checking to fine-grained component analysis, and culminating in a holistic judgment of the entire reasoning process. This approach allows us to distinguish not only correct from incorrect answers, but also robustly derived solutions from those that are superficially correct but procedurally flawed.

**Final Answer Correctness (ACCU)** Given that VisAidMath comprises mathematics problems with deterministic answers, we evaluate the correctness of the final answer using Standard Accuracy (ACCU) as a baseline metric. To accurately extract the final answer from model outputs, we follow the approach of (Lu et al., 2023) and employ GPT-

4o mini as answer extractor, as it demonstrated a success rate of 99% in preliminary experiments with 200 examples. The prompts used to construct the answer extractor are described in detail in Appendix F.6. ACCU is calculated as:

$$\text{ACCU} = \frac{1}{N} \sum_{j=1}^N a_j \quad (1)$$

where  $N$  is the total number of samples and  $a_j \in \{0, 1\}$  is the binary correctness of the final answer for sample  $j$ .

**Process-Verified Accuracy (PVA)** Recognizing that standard accuracy (ACCU) cannot distinguish valid reasoning from "lucky guesses," we introduce PVA to act as a strict reliability filter. PVA refines

Model	ACCU	PVA	SPRS	LogiR	ToolF	InfoT	GeneR
<i>Task 1: General Reasoning</i>							
VL-Cogito	48.00	24.00	11.50	5.86	8.81	8.83	7.80
Qwen2.5-VL-72B	53.17	37.39	24.50	7.42	9.22	9.39	8.73
InternVL3.5-38B	62.00	50.59	41.42	8.33	9.33	9.77	9.25
GPT-4.1	59.42	49.03	39.58	8.55	9.43	9.72	9.22
o4-mini	71.92	63.34	55.00	9.26	9.68	9.94	9.66
Doubao-Seed-1.6	78.75	68.13	55.25	9.09	9.72	9.93	9.61
<i>Task 2: Direct Visual-aided Reasoning</i>							
VL-Cogito	49.17	20.20	8.58	5.20	8.50	8.31	6.94
Qwen2.5-VL-72B	52.25	34.79	21.42	7.13	9.18	9.27	8.54
InternVL3.5-38B	63.92	20.20	30.67	7.93	9.23	9.68	9.19
GPT-4.1	62.42	52.34	44.17	8.72	9.44	9.78	9.38
o4-mini	73.00	53.84	37.92	8.29	9.53	9.44	9.23
Doubao-Seed-1.6	77.33	62.03	47.58	8.63	9.64	9.72	9.40
<i>Task 3: Indirect Visual-aided Reasoning</i>							
VL-Cogito	48.67	25.50	12.67	6.02	8.82	8.84	7.89
Qwen2.5-VL-72B	54.67	37.99	25.92	7.36	9.22	9.44	8.67
InternVL3.5-38B	60.00	47.72	38.67	8.24	9.39	9.76	9.26
GPT-4.1	60.50	51.38	44.67	8.72	9.46	9.75	9.32
o4-mini	73.50	65.02	62.08	9.31	9.74	9.94	9.65
Doubao-Seed-1.6	81.00	69.84	59.25	9.06	9.76	9.90	9.56

Table 3: Comprehensive performance evaluation of different models on three reasoning tasks. We assess multiple models on *General Reasoning*, *Direct Visual-aided Reasoning*, and *Indirect Visual-aided Reasoning*. The metrics include **ACCU** (Accuracy, in %) and our proposed **SPRS** (Solution Process Robustness Score). SPRS is composed of four sub-dimensions, each rated on a 0-10 scale: **LogiR** (Logical Rigor), **ToolF** (Tool-Free Feasibility), **InfoT** (Information Traceability), and **GeneR** (Generality of the Method). A key finding is the significant gap between accuracy (ACCU) and process robustness (SPRS) across all models, underscoring that a correct answer does not guarantee a sound and verifiable reasoning process. For each task, the best score for each metric is highlighted in **green**.

the initial accuracy score by considering a solution valid only if its final answer is correct *and* its reasoning process is judged to be holistically sound.

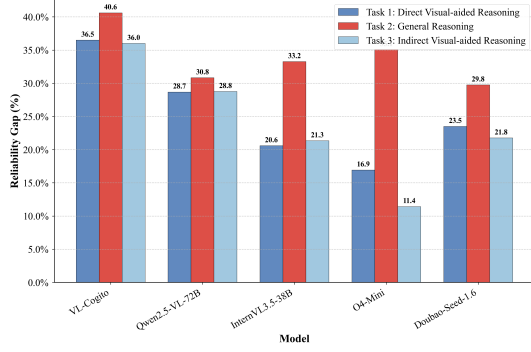
Instead of relying on a rigid, rule-based threshold, we leverage the evaluator model’s own meta-reasoning capabilities. For each solution, we employ a two-step assessment process. First, we prompt the model to provide granular scores across several evaluation dimensions ( $\mathcal{D} = \text{LogiR, ToolF, InfoT, GeneR}$ ), detailed in Appendix F.4. Second, based on its own fine-grained analysis, we prompt the model to make a holistic, binary judgment on the overall validity of the reasoning process. This yields a binary validity score,  $V_j$ , for each sample  $j$ . Formally, PVA is defined as:

$$\text{PVA} = \frac{1}{N} \sum_{j=1}^N a_j \cdot V_j \quad (2)$$

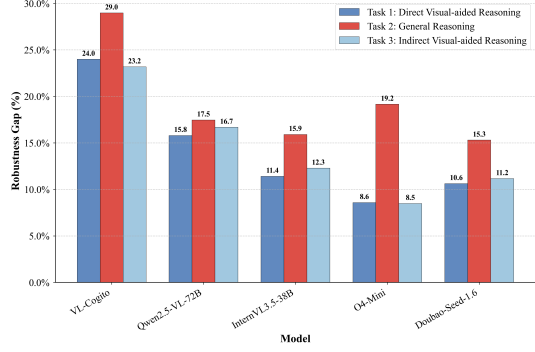
where  $a_j$  is 1 if the final answer of sample  $j$  is

correct and 0 otherwise, and  $V_j$  is the binary validity judgment (1 for valid, 0 for invalid) directly provided by the evaluator model. This approach allows the assessment to capture nuanced flaws that a simple threshold might miss, treating the LLM not just as a scorer, but as a qualitative judge of the entire reasoning chain.

**Solution Process Robustness Score (SPRS)** The framework culminates in the Solution Process Robustness Score (SPRS), which offers the most granular level of assessment. Distinct from the binary filtering of PVA, SPRS provides a continuous, fine-grained quality score for all correctly answered problems. It quantifies the overall robustness of a solution by multiplicatively aggregating its scores across all process dimensions. This mechanism, reflecting a “short-plank effect,” heavily penalizes any “weak links” in the reasoning chain. It is cal-



(a) The **Reliability Gap** (“ACCU - PVA”) across the three reasoning tasks.



(b) The **Robustness Gap** (“ACCU - SPRS”) across the three reasoning tasks.

Figure 5: Performance degradation from surface accuracy (ACCU) to process-level evaluation. The Reliability Gap (a) measures the proportion of correct answers with procedurally invalid reasoning. The Robustness Gap (b) measures the total drop in solution quality. Both gaps are most pronounced in the Direct Visual-Aided Reasoning (D-VAR) task, highlighting its unique challenge.

culated as:

$$\text{SPRS} = \frac{1}{N} \sum_{j=1}^N a_j \cdot \prod_{i \in \mathcal{D}} \left( \frac{S_{i,j}}{10} \right) \quad (3)$$

The gating by  $a_j$  ensures that only correctly answered samples contribute to the score. The product  $\prod$  over the normalized dimension scores ( $S_{i,j}/10$ ) ensures that even a single low-quality dimension significantly depresses the overall score for that sample, thus rewarding solutions that are consistently strong across all aspects of reasoning.

Together, ACCU, PVA, and SPRS form a layered diagnostic view that is particularly useful for the three-task design of VisAidMath, as it allows us to disentangle whether visual aid affects final correctness, process validity, or the overall robustness of mathematical reasoning.

### 3.3 Robustness of LLM-as-a-Judge Process Evaluation

Since PVA and SPRS rely on LLM-as-a-judge rather than human annotation, their credibility depends on the stability and consistency of the evaluator. To examine this issue, we conduct dedicated robustness analyses from two perspectives: (1) stability under sampling perturbations for repeated scoring by the same judge, and (2) consistency across different judge models. In all experiments, GPT-4.1 is used as the primary evaluator in the main benchmark, and detailed statistics are provided in Appendix G.

**Stability under Sampling Perturbations.** We repeatedly score the same set of 100 randomly sam-

pled solution processes while varying the evaluator’s decoding parameters, with temperature in  $\{0.4, 0.6, 0.8\}$  and top- $p$  in  $\{0.4, 0.7, 1.0\}$ . For each evaluation dimension and the aggregated SPRS, we report the mean  $\mu$ , standard deviation  $\sigma$ , coefficient of variation (CV), and Spearman correlation  $\rho$  with the mean vector of repeated reviews. GPT-4.1 exhibits low relative variation across dimensions (CV generally in the range of 0.10-0.20) and high rank-order consistency ( $\rho = 0.84$ -0.94). The aggregated SPRS also remains stable, achieving a high Spearman correlation of 0.94. These results indicate that the proposed process evaluation is not overly sensitive to moderate sampling perturbations and preserves reliable relative rankings among solutions.

**Cross-Judge Consistency.** To further assess whether the evaluation trends are specific to a single judge, we compare GPT-4.1 with Qwen3-Max on the same solution set. The two judges show strong agreement in relative scoring trends, with Pearson correlations of 0.72-0.80 across the four dimensions and 0.80 for SPRS. At the same time, Qwen3-Max tends to assign lower absolute scores than GPT-4.1, indicating a systematic difference in scoring scale rather than a disagreement in comparative judgments. Overall, these robustness analyses support the use of GPT-4.1 as a stable judge for process-level evaluation in VisAidMath.

### 3.4 Main Results

#### Initial Performance via Standard Accuracy

The complete performance results are presented in Tables 31, 2, and 32, with a summary of lead-

ing models in Table 3. On the metric of standard accuracy (ACCU), top-performing models demonstrate high scores. For instance, Doubao-Seed-1.6 achieves a peak ACCU of **81.00%** on the Indirect Visual-Aided Reasoning task and **77.33%** on the Direct Visual-Aided Reasoning task. While these figures indicate a strong capability for arriving at correct final answers, they alone do not validate the underlying reasoning processes used to obtain them. This task-by-metric comparison is essential for interpreting the role of visual aid, because it shows that the current bottleneck lies less in the usefulness of visual representations themselves than in models’ ability to generate and exploit them reliably.

### Performance Collapse under Deeper Scrutiny

The insufficiency of ACCU as a standalone metric becomes evident when our Funnel Evaluation Framework is applied. As shown in Table 3, a significant and universal decline is observed from ACCU to both PVA and SPRS. On the Direct Visual-Aided Reasoning task, for example, Doubao-Seed-1.6’s 77.33% ACCU declines to a PVA of 62.03% and an SPRS of 47.58%. We define this stark discrepancy between apparent correctness and procedural soundness as the “**reasoning illusion**”. This phenomenon provides powerful evidence that VisAidMath effectively probes deep reasoning deficiencies that are masked by conventional accuracy-only evaluations. A quantitative analysis of this illusion and its implications for the VisAidMath benchmark is the central focus of Section 4.

## 4 Analysis

Following the discovery of the universal “reasoning illusion”, this chapter provides a deep analysis of this phenomenon. We first offer quantitative proof demonstrating how VisAidMath uniquely exposes model deficiencies, and then present a qualitative diagnosis to understand the root causes behind these failures.

### 4.1 Reasoning Gap Quantified

**Reliability and Robustness Gaps** To objectively measure the performance drop from surface accuracy to procedural quality, we define two metrics: the **Reliability Gap** (“ACCU - PVA”) and the **Robustness Gap** (“ACCU - SPRS”). The former quantifies the proportion of correct answers derived from flawed processes, while the latter measures

the overall decline in solution quality. Figure 5 visualizes these two gaps across our three reasoning tasks.

The analysis reveals a critical finding: the **Direct Visual-Aided Reasoning (D-VAR) task induces a dramatically larger gap** in both reliability and robustness compared to the other tasks. This indicates that while models may seem competent on the surface, their reasoning processes are particularly fragile when required to directly engage with visual information. For instance, InternVL3.5-38B on the D-VAR task exhibits a massive 43.7-point Reliability Gap and a 33.2-point Robustness Gap. This provides strong quantitative evidence that the core task of VisAidMath is uniquely effective at pressuring models to reveal their underlying reasoning deficiencies.

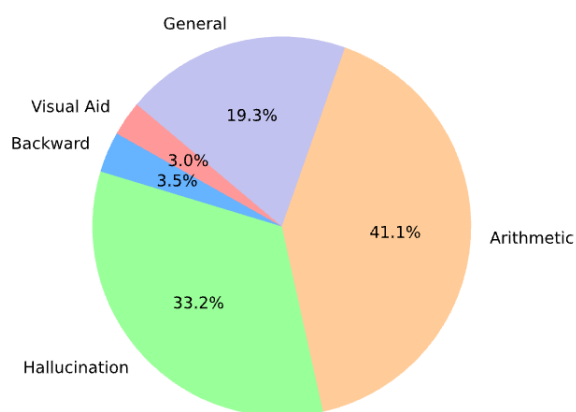
**Diagnosing Failure Patterns** The interplay between PVA and SPRS also serves as a powerful diagnostic tool for identifying distinct model failure modes. The relationship between these two metrics is not uniform across models, revealing different behavioral patterns:

- **Case A: Catastrophic Failures (SPRS > PVA).** For models like InternVL3.5-38B on the D-VAR task (PVA: 20.20%, SPRS: 30.67%), the SPRS score is notably higher than the PVA score. This pattern suggests the model frequently makes critical, “all-or-nothing” errors that cause its solutions to be entirely invalidated by the PVA filter.
- **Case B: Systemic Minor Flaws (PVA > SPRS).** In contrast, models like Doubao-Seed-1.6 (PVA: 62.03%, SPRS: 47.58%) exhibit a higher PVA than SPRS. This indicates the model is more adept at avoiding fatal errors, but its procedurally “valid” solutions are often rife with minor inaccuracies, which are penalized by the SPRS, dragging down the overall robustness score.

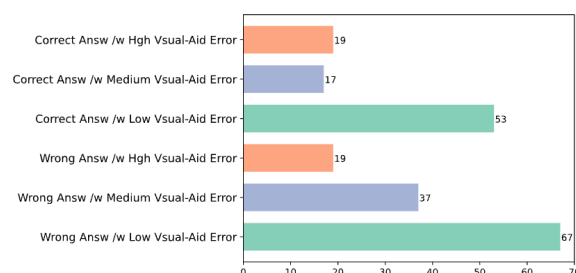
This ability to distinguish between models prone to catastrophic accidents and those exhibiting systemic sloppiness underscores the diagnostic depth of our framework.

### 4.2 Qualitative Diagnosis

**Evasion of Visual Reasoning** A primary cause for the reasoning gap is that models strongly tend to *evoke* the intended visual-aided path. Our manual



(a) Distribution of reasoning strategies for correctly solved problems. The minimal use of "Visual-Aided" reasoning (3.0%) reveals a strong model tendency to evade the intended solution path.



(b) Correlation between visual-aid error severity and final answer correctness. High-severity errors are strongly predictive of an incorrect final answer.

Figure 6: Qualitative diagnosis of the reasoning gap.

analysis of 200 correctly answered D-VAR samples shows that a staggering majority relied on non-visual shortcuts (Figure 6a). Only 3.0% of correct solutions were achieved by generating and using visual aids as intended. The dominant strategies were pure arithmetic (41.1%) or general reasoning that ignored the visual context (19.3%). This circumvention of the core task is a key factor explaining how high ACCU scores can mask poor underlying reasoning processes.

**The High Cost of Flawed Attempts** Furthermore, when models do attempt the visual reasoning path, errors in this intermediate step are often fatal to the final outcome. As demonstrated in Figure 6b, there is a strong correlation between the severity of visual-aid errors and the correctness of the final answer. Solutions with significant visual-aid errors are overwhelmingly more likely to result in an incorrect final answer. This high cost of flawed attempts further explains the significant performance gaps observed in the D-VAR task, as engaging with the task incorrectly is highly detrimental. A more detailed breakdown of these error types can be found in Appendix 28b.

**Format Compliance vs. Spatial Reasoning Failure.** The main bottleneck in D-VAR lies in spatial reasoning rather than in adhering to the required output format. The format specification is intentionally lightweight and explicitly scaffolded in the prompt, requiring only a simple structured description of the constructed visual aid. Nevertheless, even strong instruction-following models exhibit

substantial drops on D-VAR under process-level evaluation, which suggests that the observed difficulty cannot be attributed primarily to formatting. Qualitative inspection further shows that the dominant errors are semantic rather than syntactic: models generate geometrically impossible or task-irrelevant auxiliary constructions, refer to non-existent points, or bypass the intended construction step and revert to pure textual or arithmetic reasoning. Together, these patterns indicate that D-VAR primarily probes whether a model can construct and use spatially meaningful auxiliary representations, rather than whether it can satisfy a superficial output template.

## 5 Conclusion

In this paper, we lay the groundwork for mathematical problem solving using multi-modal reasoning steps. We introduce VisAidMath, a benchmark designed to investigate the visual-aided reasoning capabilities of both large language models and large multi-modal models. Experiments on mainstream models demonstrate deficiencies in deducing visual aids and the corresponding textual reasoning steps. We also conducted fine-grained quantitative and qualitative analyses to reveal disparities in visual-aid reasoning. This exposes a divide between passive perception and active, visually grounded deduction in current LMMs. VisAidMath thus stands as both a challenging benchmark and a guiding paradigm for advancing visual grounded reasoning.

## Limitations

The limitation of **VisAidMath** is three-folded. First, dataset is restricted to 1200 samples since both collection, annotation and verification of mathematical problems acquire heavy manual work to satisfy dataset principles. Such mathematical problems with visual aids cost more human efforts to understand each segment before judgment. Secondly, deficiency of mainstream machine translation systems in mathematical domain could introduce various translation errors, thus enhancing complexity for problem solving and subsequent evaluation. Thirdly, we cannot conduct comprehensive analysis of visual-aided reasoning with image generation, since current LMMs remain significant deficiency in mathematical image generation. No negative social impact will be provided from our math-centric work, expecting only to enhance further understanding of LLM reasoning.

## Acknowledgement

This work was supported in part by the Science and Technology Development Fund of Macau SAR (Grant Nos. FDCT/0007/2024/AKP, EF2024-00185-FST), the UM and UMDP (Grant Nos. MYRG-GRG2024-00165-FST-UMDF, MYRG-GRG2025-00236-FST), the Tencent AI Lab Rhino-Bird Research Program (Grant No. EF2023-00151-FST), the Stanley Ho Medical Development Foundation (Grant No. SHMDF-AI/2026/001), and the National Natural Science Foundation of China (Grant No. 62266013).

## References

- AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Guanhua Chen, Runzhe Zhan, Derek F. Wong, and Lidia S. Chao. 2023. **Multi-level curriculum learning for multi-turn dialogue generation**. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:3958–3967.
- Guanhua Chen, Runzhe Zhan, Derek F. Wong, and Lidia S. Chao. 2024. **Dynamic curriculum learning for conversation response selection**. *Knowl. Based Syst.*, 293:111687.
- Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022. Uni-geo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746*.
- Yuri Chervonyi, Trieu H Trinh, Miroslav Oľšák, Xiaomeng Yang, Hoang Nguyen, Marcelo Menegali, Junehyuk Jung, Vikas Verma, Quoc V Le, and Thang Luong. 2025. Gold-medalist performance in solving olympiad geometry with alpheageometry2. *arXiv preprint arXiv:2502.03544*.
- Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.
- Daocheng Fu, Zijun Chen, Renqiu Xia, Qi Liu, Yuan Feng, Hongbin Zhou, Renrui Zhang, Shiyang Feng, Peng Gao, Junchi Yan, et al. 2025a. Trustgeogen: Scalable and formal-verified data engine for trustworthy multi-modal geometric problem solving. *arXiv preprint arXiv:2504.15780*.
- Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. 2025b. Refocus: Visual editing as a chain of thought for structured image understanding. *arXiv preprint arXiv:2501.05452*.
- Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR.
- Google. 2023. **Google gemini pro vision**. In *Google*. Google.
- Xinyu Guan, Li Lina Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. **rstar-math: Small LLMs can master math reasoning with self-evolved deep thinking**. In *Forty-second International Conference on Machine Learning*.
- Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. 2025a. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*.
- Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Rui Huang, Haoquan Zhang, Manyuan Zhang, Jiaming Liu, Shanghang Zhang, Peng Gao, et al. 2025b. Can we generate images with cot? let’s verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*.

- Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14953–14962.
- 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. *arXiv preprint arXiv:2103.03874*.
- Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2022. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*.
- Gueyoung Jung, Matti A Hiltunen, Kaustubh R Joshi, Richard D Schlichting, and Calton Pu. 2010. Mistral: Dynamically managing power, performance, and adaptation cost in cloud infrastructures. In *2010 IEEE 30th International Conference on Distributed Computing Systems*, pages 62–73. IEEE.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- larryflynt. 2023. <https://github.com/larryflynt/image-concat>. In *Github*.
- Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857.
- Can Li, Ying Liu, Ting Zhang, Mei Wang, and Hua Huang. 2025. Visiomath: Benchmarking figure-based mathematical reasoning in llms. *arXiv preprint arXiv:2506.06727*.
- Adam Dahlgren Lindström and Savitha Sam Abraham. 2022. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*.
- Dairu Liu, Ziyue Wang, Minyuan Ruan, Fuwen Luo, Chi Chen, Peng Li, and Yang Liu. 2025. Visual abstract thinking empowers multimodal reasoning. *arXiv preprint arXiv:2505.20164*.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- 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*.
- Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*.
- Jingkun Ma, Runzhe Zhan, and Derek F. Wong. 2023. **Yu sheng: Human-in-loop classical Chinese poetry generation system**. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 57–66, Dubrovnik, Croatia. Association for Computational Linguistics.
- Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.
- OpenAI. 2023a. **Gpt-3.5-turbo**. In *OpenAI*.
- OpenAI. 2023b. **Gpt-4**. In *OpenAI*. OpenAI.
- OpenAI. 2023c. **Gpt-4.1**. In *OpenAI*. OpenAI.
- OpenAI. 2023d. **Gpt-4v**. In *OpenAI*. OpenAI.
- OpenAI. 2023e. **o4-mini**. In *OpenAI*. OpenAI.
- Jaewoo Park, Jungyang Park, Dongju Jang, Jiwan Chung, Byungwoo Yoo, Jaewoo Shin, Seonjoon Park, Taehyeong Kim, and Youngjae Yu. 2025. Explain with visual keypoints like a real mentor! a benchmark for multimodal solution explanation. *arXiv preprint arXiv:2504.03197*.
- Shuai Peng, Di Fu, Liangcai Gao, Xiuqin Zhong, Hongguang Fu, and Zhi Tang. 2024. Multimath: Bridging visual and mathematical reasoning for large language models. *arXiv preprint arXiv:2409.00147*.
- Ji Qi, Ming Ding, Weihang Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, et al. 2024. Cogcom: A visual language model with chain-of-manipulations reasoning. *arXiv preprint arXiv:2402.04236*.
- Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

- Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. 2024. [Math-LLaVA: Bootstrapping mathematical reasoning for multimodal large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4663–4680, Miami, Florida, USA. Association for Computational Linguistics.
- Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. 2025. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*.
- Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.
- Shichao Weng, Zhiqiang Wang, Yuhua Zhou, Rui Lu, Ting Liu, Zhiyang Teng, Xiaozhang Liu, and Hanmeng Liu. 2025. Geosketch: A neural-symbolic approach to geometric multimodal reasoning with auxiliary line construction and affine transformation. *arXiv preprint arXiv:2509.22460*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Qinzhao Wu, Qi Zhang, Jinlan Fu, and Xuan-Jing Huang. 2020. A knowledge-aware sequence-to-tree network for math word problem solving. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7137–7146.
- Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. 2024. Mind’s eye of llms: visualization-of-thought elicits spatial reasoning in large language models. *Advances in Neural Information Processing Systems*, 37:90277–90317.
- Yuhuai Wu, Markus N Rabe, Wenda Li, Jimmy Ba, Roger B Grosse, and Christian Szegedy. 2021. Lime: Learning inductive bias for primitives of mathematical reasoning. In *International Conference on Machine Learning*, pages 11251–11262. PMLR.
- Ruifeng Yuan, Chenghao Xiao, Sicong Leng, Jianyu Wang, Long Li, Weiwen Xu, Hou Pong Chan, Deli Zhao, Tingyang Xu, Zhongyu Wei, et al. 2025. VI-cogito: Progressive curriculum reinforcement learning for advanced multimodal reasoning. *arXiv preprint arXiv:2507.22607*.
- Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. 2023. How well do large language models perform in arithmetic tasks? *arXiv preprint arXiv:2304.02015*.
- Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*.
- Yiming Zhang, Zicheng Zhang, Xinyi Wei, Xiaohong Liu, Guangtao Zhai, and Xiongkuo Min. 2025. [Illusionbench: A large-scale and comprehensive benchmark for visual illusion understanding in vision-language models](#). In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.
- Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. 2025. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1702–1713.

<b>Contents</b>	
<b>A Future Direction</b>	<b>13</b>
<b>B Dataset Preparation</b>	<b>14</b>
B.1 Machine Translation . . . . .	14
B.2 Data Processing . . . . .	14
B.3 Formalization . . . . .	14
<b>C Dataset Analysis</b>	<b>14</b>
C.1 Metadata . . . . .	14
C.2 Data Source . . . . .	14
<b>D Examples for Different Categorizations</b>	<b>15</b>
D.1 Math Branch . . . . .	15
D.2 Visual Aid Type . . . . .	15
D.3 Complexity . . . . .	15
<b>E Dataset Collection Detail</b>	<b>15</b>
E.1 Annotation Details . . . . .	15
E.2 Caption Writing Templates . . . . .	17
E.3 Annotation Roles . . . . .	17
E.4 Dataset Creation Pipeline . . . . .	18
E.5 Human Annotation Interface . . . . .	18
<b>F Detail Experiment Settings</b>	<b>25</b>
F.1 Hyperparameters . . . . .	25
F.2 Reasoning Prompter . . . . .	25
F.3 Instruction across Answer Types . . . . .	25
F.4 Detail of Evaluation Dimensions . . . . .	25
F.5 Visual Aid Extraction . . . . .	25
F.6 Answer Extraction Prompter . . . . .	25
<b>G Robustness Analysis of LLM-as-a-Judge Evaluation</b>	<b>25</b>
G.1 Experimental Setup . . . . .	25
G.2 Repeated-Scoring Stability . . . . .	26
G.3 Cross-Judge Consistency . . . . .	27
<b>H More Experimental Results</b>	<b>28</b>
H.1 Results of other tasks . . . . .	28
H.2 Quantitative Analysis . . . . .	28
H.2.1 Performance across Math Branches . . . . .	28
H.2.2 Performance across Complexities . . . . .	32
H.2.3 Performance across Visual Aids . . . . .	32
H.3 N-gram Similarities . . . . .	33
<b>I In-depth Model Analysis</b>	<b>37</b>
I.1 General Reasoning Tendency . . . . .	37
I.2 Visual Aid Inference Capability . . . . .	37

<b>J Supplementary Quantitative Analysis</b>	<b>37</b>
<b>K Related Work</b>	<b>38</b>

## Appendix

### A Future Direction

**Spatial Capability** Although recent progress in mathematical reasoning has been strongly driven by larger datasets and stronger training recipes, most mainstream work still remains concentrated on textual inference. Our results show that current LLMs and LMMs are still weak at producing explicit visual reasoning steps, revealing limitations in spatial understanding, mental construction, and visually grounded planning. We believe that two research directions are especially important and complementary. In the short term, better data construction, supervision design, and training strategies may provide the most immediate improvements. In the longer term, however, more fundamental advances may require architectural innovations that explicitly support spatial reasoning and tighter integration between textual and visual inference.

**Mathematical Image Generation** Preliminary experiments find mainstream models exhibit poor mathematical image generation performance, thus further captioning each mathematical images to explore visual reasoning step inference. Primary model deficiencies fall in: mathematical image caption comprehension, spatial relationships apprehension, lack of numerical precision, significant stylization discrepancies in the images, and more. Generate image drawing code can increase the overall drawing precision, while suffering from plenty of code errors. There lies a long research road in mathematical image generation before fully exploration of textual-visual interconnected inference.

**Evaluation Metrics** Reasoning non-uniqueness enhances evaluation complexity of visual aids generation. Different viewing angle, relative element size, and styles can alter perceptual features instead of semantic feature. Visual-aid can be captioned by multiple correct expressions with semantic remains stable. Therefore, future evaluation metrics research for visual-aid should be directed toward semantic-based method.

## B Dataset Preparation

### B.1 Machine Translation

Since most of our data and their corresponding captions were in Chinese, we first translated all text into English. Open-source machine-translation (MT) models often exhibit deficiencies in semantic fidelity and numerical accuracy within mathematical contexts. Through manual sampling and comparison, we identified Baidu Translate<sup>3</sup> and DeepL<sup>4</sup> as high-quality services for our specific needs. Therefore, we employed both platforms, splitting each sample’s translation between the two to mitigate potential data leakage.

Importantly, our English benchmark is not constructed through machine translation alone. Instead, the translation pipeline follows three stages: Chinese source text, preliminary machine translation, and human review and refinement. In the human refinement stage, annotators revise translation errors, improve naturalness and fluency, and verify that the mathematical logic of the original problem is preserved in the final English version.

To further validate the quality of the translated benchmark after human refinement, we conducted an additional human evaluation on the final English data. We randomly sampled 50 representative items from the dataset and invited two bilingual evaluators, both Chinese-native master’s students trained in an English academic environment and independent of the original annotation process, to perform blind back-to-back assessments. We focus on **Logical Equivalence**, which measures whether the final English version preserves the mathematical meaning and reasoning structure of the original Chinese problem without substantive deviation, using a 5-point scale. The average score is **3.52/5**, indicating that the translated data generally preserve the original mathematical content with acceptable quality for research use.

### B.2 Data Processing

We further process the annotated data to match the model-input format and the requirements of our evaluation protocol. Because many models accept only a single image per generation round, we use the tool described in (larryflynt, 2023) to merge either multiple visual-context images into one when necessary. Decimal answers are rounded to three

decimal places, and fractional answers are rewritten in the form “numerator/denominator.” Problems with free-form answers are reformulated as multiple-choice or true/false questions, with the correct option marked as the answer. Finally, we perform a manual validation pass, revising the captions of both visual-context and visual-aid images to ensure completeness and accuracy.

### B.3 Formalization

Each sample is stored in its own directory containing two subfolders—one for visual-context images and one for visual-aid images—and a `data.json` file. Images are named in the order in which they appear in the problem or rationale (e.g., `1.png`, `2.png`). Any merged image is saved as `concatenate.png`. The `data.json` file holds all text and metadata, including the original (untranslated) text when available. The dataset is released under the CC-BY-SA-4.0 license, and VisAidMath is intended for research use only.

## C Dataset Analysis

### C.1 Metadata

We list the manually annotated metadata for each sample in Table 4. “visual context” is optional depending on whether image is provided along with the problem. “choices” is not empty when question form belongs to multiple choices or true/false. “language” stamp the original language of each problem. We also record the Chinese version text before machine translation with “zh\_” prefix at the bottom of each data sample.

### C.2 Data Source

In accordance with the aforementioned principles, the VisAidMath benchmark has been manually collected and annotated using a diverse and balanced set of data sources. Through an extensive search and careful examination on a case-by-case basis, we discovered that the Chinese community offers a larger pool of mathematical problems with visual aids across various complexity levels and mathematical branches compared to other communities. As a result, we primarily collected data from Chinese sources and subsequently performed machine translation. To ensure formula consistency, we replace LaTeX formulas with placeholders before translation and refill afterward. We also provide human validation of all samples to further ensure translation quality. To ensure a range of difficulty

<sup>3</sup><https://fanyi.baidu.com/>

<sup>4</sup><https://www.deepl.com/translator>

Category	Detail
question	Text of mathematical problem
visual context: image path	Relative path of visual context image
visual context: caption	Caption of visual context image
visual aid: image path	Relative path of visual aids image
visual aid: caption	Caption of visual aids image
choices	Options for multiple choice problems
question form	Question form includes: multiple choice, true/false, free form
answer form	Integer, decimal, fraction, and choice (for multiple choice and true/false problems)
answer	Answer of this mathematical problem
metadata: language	Original language of this mathematical problem.
metadata: source	Data source
metadata: math branch	Mathematical branch
metadata: drawing type	type of visual aids

Table 4: Detail metadata for **VisAidMath**

levels, we categorized the data samples based on their sources into the following categories: Easy, Medium and Hard. Additionally, metadata has been included for further in-depth analysis, discussed in section C.1.

The complete complexity categorization of each source is listed in 5. Particularly, since “AP Calculus” consists of both easy and medium level mathematical problems, we consider questions satisfying one of the following conditions as medium level: 1) involve coordinate axes rotation 2) cannot be resolved in one step leveraging Newton-Leibniz formula.

## D Examples for Different Categorizations

### D.1 Math Branch

As shown in Table 6, 7, 8 and 9, this section presents examples of mathematical problems from different branches: plane geometry, solid geometry, analytic geometry, and calculus. Each example includes a question and visual aids to help illustrate the concepts.

### D.2 Visual Aid Type

### D.3 Complexity

## E Dataset Collection Detail

### E.1 Annotation Details

The quality of annotation plays a crucial role in ensuring the reliability of the benchmark, and the annotation of metadata significantly impacts the depth

and breadth of analysis. In the annotation process, we instructed the annotators to label various metadata, including the *math category*, *source*, *visual aids’ type*, *question form*, *choices*, *answer form*, *answer*, and *language*. Given that the formalization of mathematical problems can vary significantly across different data sources, we specifically asked the annotators to manually extract the question and answer text from the  $\text{\LaTeX}$  file to ensure completeness and accuracy. Following the approach used in MathVista (Lu et al., 2023), we transformed free-form questions without a purely numerical answer into multiple-choice or True/False questions. This deterministic transformation allows for a more robust evaluation. To create precise and comprehensive descriptions of the visual context and visual aids, we provided the annotators with caption writing templates that were designed to simplify the complexity of caption writing. For more details, please refer to section E.2. Four Chinese undergraduate students who majored in STEM during their high school period were carefully selected to form our annotation team. The collectors and annotators were compensated salary aligns with average price in local data crowdsourcing market for each data sample they collected or annotated. Furthermore, a graduate student specializing in mathematics was appointed as the verifier, offering professional annotation guidance, validation, and refinement throughout the process. All participants were volunteers who provided informed consent.

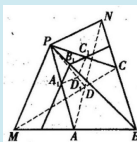
Data Source	Detail
High Textbook	Chinese high school textbook
Middle Practice	Chinese high school practice sheet
AP Easy	AP calculus (categorized into Easy category)
Middle Simulate	Chinese middle school simulated examination
AP Middle	AP calculus (categorized into Medium category)
High Practice	Chinese high school practice sheet
DSE Final	HKDSE final examination
High Final	Chinese high school final examination
High Simulate	Chinese high school simulated examination
Math Analysis Demidovich Textbook	Demidovich Problems in Mathematical Analysis
Analytic Geometry Lv Textbook	Analytic geometry textbook written by Lingen Lv
CMO Final	Chinese Mathematical Olympiad
CMO Practice	Chinese Mathematical Olympiad practice sheet
AIME Final	American Invitational Mathematics Examination (AIME)
AMC 8 Practice	American Mathematics Competition 8 (AMC 8)
AMC 10 Final	American Mathematics Competition 10 (AMC 10)

Table 5: Detail of data sources

**ID:** 1114

**Question:** determine that the sum of the squares of the lengths of the three medians of a triangle is equal to the sum of the squares of the lengths of the three sides  $\frac{3}{4}$ .

**Visual Aids:**

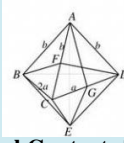


**Caption of Visual Aids:** Let the three vertices of a triangle be the midpoint of  $A, B, C, AB$ , the midpoint of  $F, AC$ , the midpoint of  $E, BC$ , and the midpoint of  $D$ , connecting  $AD, BE, CF$

Table 6: Example of mathematical problems of plane geometry.

**ID:** 1006

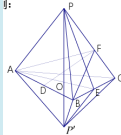
**Visual Context:**



**Caption of Visual Context:** Two congruent triangles are glued together to obtain a hexahedron with all the dihedral angles equal, and the length of the shortest prong of the hexahedron is 2.

**Question:** Given that two congruent triangular pheons are glued together to obtain a hexahedron with all the dihedral angles equal, and that the shortest prong of the hexahedron is 2, the distance between the two farthest vertices is

**Visual Aids:**

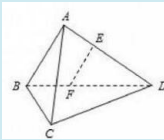


**Caption of Visual Aids:** Make a plane  $BDF \perp$  with edges  $AC$  intersecting  $AC$  at  $F$

Table 7: Example of mathematical problems of solid geometry.

**ID:** 619

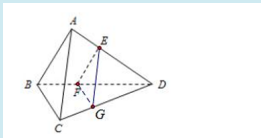
**Visual Context:**



**Caption of Visual Context:** In the trigonometry  $A - BCD$ ,  $AB \perp AD, BC \perp BD$ , the plane  $ABD \perp$  the plane  $BCD$ , points  $E, F$  ( $E$  do not coincide with  $A, D$ ) are on the prisms  $AD, BD$  and  $EF \perp AD$ .  $EF // \text{Plane } ABC$

**Question:** In the trigonal pheasant  $A - BCD$ ,  $AB \perp AD, BC \perp BD$ , plane  $ABD \perp$  plane  $BCD$ , points  $E, F$  ( $E$  and  $A, D$  do not coincide) on the prong  $AD, BD$  respectively, and  $EF \perp AD, EF // \text{plane } ABC$  determine whether  $AD \perp AC$ .

**Visual Aids:**



**Caption of Visual Aids:** Take a point  $G$  on the line segment  $CD$ , connect  $FG, EG$  so that  $FG // BC$

Table 8: Example of mathematical problems of analytic geometry.

The task was described as minimal risk.

## E.2 Caption Writing Templates

Mathematical graphs are consists of shapes and elements bound with specific relation or theorem. To reduce manual annotation work and enhance caption consistency, we standardize the caption writing for visual context and visual aids by defining templates for certain visual elements. The annotators should caption image referring to these templates as listed in 19.

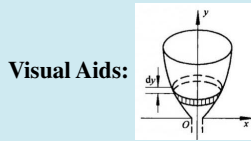
## E.3 Annotation Roles

The dataset creation pipeline involves four key roles:

- **Administrator:** This role assigns daily collection tasks based on the progress and previous annotation feedback.
- **Collector:** The collector searches for data that satisfies the assigned collection tasks. The collected data should be in PDF format and later transformed into  $\text{\LaTeX}$  files using OCR.

**ID:** 939

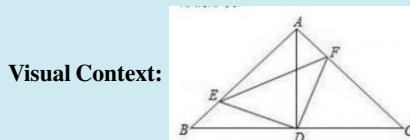
**Question:** Can the surface of a container formed by rotating the curve  $y = Cx^4$  about the vertical axis  $Oy$  cause the liquid surface to fall uniformly as it flows from the bottom of the container?



**Caption of Visual Aids:** In a geometric body obtained by rotating an upward parabola along the  $y$ -axis,  $dx$  is a line segment parallel to the  $x$ -axis, and the volume enclosed by it is drawn

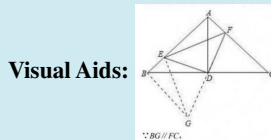
Table 9: Example of mathematical problems of calculus and functions.

**ID:** 10



**Caption of Visual Context:** In  $\text{Rt } \triangle ABC$ ,  $\angle BAC = 90^\circ$ , point  $D$  is the midpoint of  $BC$ , points  $E, F$  are points on  $AB, AC$  respectively, and  $ED \perp FD$ .

**Question:** As shown in the figure, in  $\text{Rt } \triangle ABC$ ,  $\angle BAC = 90^\circ$ , the point  $D$  is the midpoint of  $BC$ , the point  $E, F$  is the point on  $AB, AC$ , and  $ED \perp FD$ . Can a triangle be formed with the segment  $BE, EF, FC$  as its side? If so, determine the shape of the triangle.



**Caption of Visual Aids:** Make  $BG \parallel FC$ , intersect with  $FD$  extension line at  $G$ , connect  $EG$

Table 10: Example of mathematical problem with auxiliary line as visual aid.

- **Annotator:** The annotator first validates and refines the  $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$  files by comparing the original PDF files provided by the collector with the transformed  $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$  files. Then, the annotator performs interactive labeling using our designed tool. To ensure a balanced distribution across different categories, the annotator regularly analyzes the data distribution and provides feedback on the current progress and any annotation issues to the collector and administrator.
- **Verifier:** The verifier is responsible for validating the categorization and data quality. If labels are not appropriate, they adjust the annotated captions of the context and visual aids.

#### E.4 Dataset Creation Pipeline

The pipeline is shown in the Figure 7.

#### E.5 Human Annotation Interface

Shown in the Figure 8.

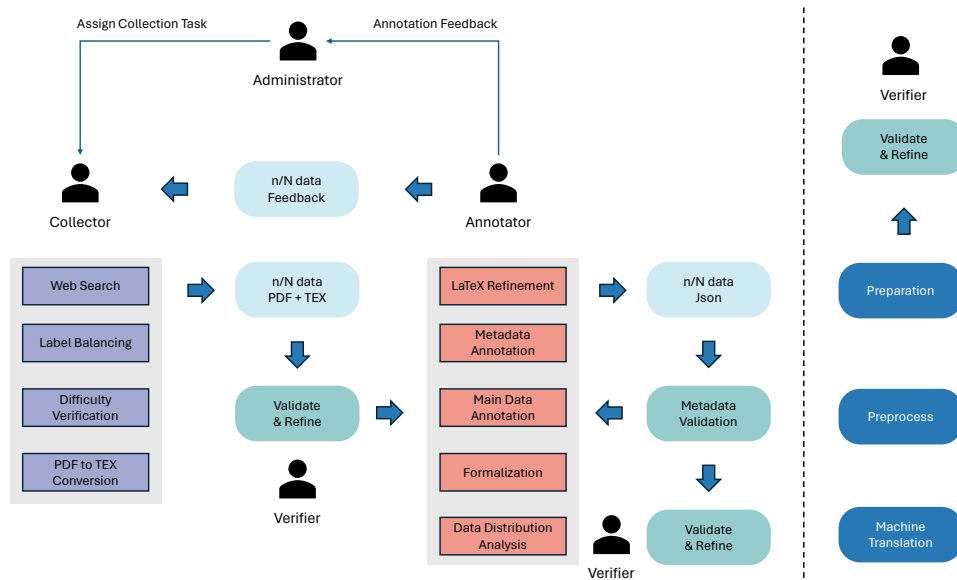
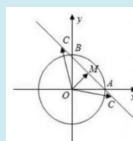


Figure 7: Pipeline involving data collection, annotation and verification.

ID: 719

**Question:** Given that the vector  $\vec{OA}, \vec{OB}$  satisfies  $|\vec{OA}| = |\vec{OB}| = 1, \vec{OA} \perp \vec{OB}, \vec{OC} = \lambda\vec{OA} + \mu\vec{OB}(\lambda, \mu \in R)$ , if  $M$  is the midpoint of  $AB$  and  $|\vec{MC}| = 1$  then the maximum value of  $\lambda + \mu$  is ( )

**Choices:** A.  $1 - \sqrt{3}$  B.  $1 + \sqrt{2}$  C.  $\sqrt{5}$  D.  $1 + \sqrt{3}$  **Visual Aids:**



**Caption of Visual Aids:** Establish a Cartesian coordinate system Draw the vector  $\vec{OA}, \vec{OB}$  M is the midpoint of AB

Table 11: Example of mathematical problem with rectangular coordinate system as visual aid.

```

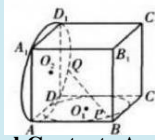
$ bash scripts/interactive_label.sh
Please enter the desired data range for annotation (for multiple questions: '1-70', for a single question: '70-70'): 1-1200
===For the specified category of labels, please enter the label numbers (e.g., (1, 2, 3)) for annotation===
=====Question 1=====
-----Please enter the label numbers. You can enter multiple labels, for example: 1-2-4. (When entering multiple options, please note which categories can include multiple labels)-----
question_form:
0. multi_choice
1. true_false
2. free_form
index: 2
-----Please enter the label numbers. You can enter multiple labels, for example: 1-2-4. (When entering multiple options, please note which categories can include multiple labels)-----
answer_form:
0. integer
1. decimal
2. fraction
3. choice
index: 1
-----Please enter the label numbers. You can enter multiple labels, for example: 1-2-4. (When entering multiple options, please note which categories can include multiple labels)-----
language:
0. chinese
1. english
2. q_en_a_ch
3. q_ch_a_en
index: 1
-----Please enter the label numbers. You can enter multiple labels, for example: 1-2-4. (When entering multiple options, please note which categories can include multiple labels)-----
source:
0. middle_final
1. middle_practice
2. middle_simulate

```

Figure 8: Annotation tool for interactive labeling

**ID:** 1099

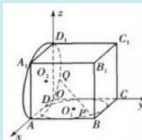
**Visual Context:**



**Caption of Visual Context:** A square  $ABCD - A_1B_1C_1D_1$  has prism length 1,  $\odot O_1$  is the incircle of the square  $ABCD$ ,  $\odot O_2$  is the outer circle of the square  $ADD_1A_1$ , and  $P, Q$  are the points on  $\odot O_1, \odot O_2$  respectively.

**Question:** As shown in the figure, the prisms of the square  $ABCD - A_1B_1C_1D_1$  have the lengths 1,  $\odot O_1$  as the tangent circle of the square  $ABCD$ ,  $\odot O_2$  as the outer circle of the square  $ADD_1A_1$ , and  $P, Q$  as the point on  $\odot O_1, \odot O_2$ . Find the range of the length of  $PQ$ .

**Visual Aids:**



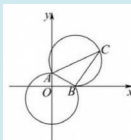
**Caption of Visual Aids:** Let the origin  $O$  and  $D$  points coincide, along  $DD_1$  Establish the  $Z$  axis in the  $DC$  direction, establish the  $y$  axis in the  $DA$  direction, and establish the  $X$  axis in the  $DC$  direction

Table 12: Example of mathematical problem with rectangular three-dimensional coordinate system as visual aid.

**ID:** 723

**Question:** Given that the plane vector  $\vec{a}, \vec{b}, \vec{c}$  satisfies the angle between  $|\vec{a}| = 1, |\vec{b}| = \sqrt{3}, \vec{a} \cdot \vec{b} = 0, \vec{c} - \vec{a}$  and  $\vec{c} - \vec{b}$  is  $\frac{\pi}{6}$ , then the maximum value of  $\vec{c} \cdot (\vec{b} - \vec{a})$  is

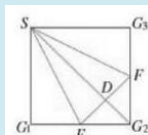
**Visual Aids:**



**Caption of Visual Aids:** Make the tangent between circle  $o$  and circle  $c$   $I_1, I_2, I_3$

Table 13: Example of mathematical problem with geometry graph as visual aid.

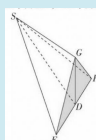
**ID:** 220



**Visual Context:**

**Caption of Visual Context:** In the square  $SG_1G_2G_3$ ,  $E, F$  are the midpoints of  $G_1G_2, G_2G_3$  respectively, and  $D$  is the midpoint of  $EF$ , connecting  $EF, SE, SF, SG_2$ .

**Question:** As shown in the figure, in the square  $SG_1G_2G_3$ ,  $E, F$  is the midpoint of  $G_1G_2, G_2G_3$  and  $D$  is the midpoint of  $EF$  respectively. If this square is folded into a tetrahedron along  $SE, SF$  and  $EF$  so that the three points of  $G_1, G_2, G_3$  coincide, and the point of coincidence is labeled  $G$ , then which of the edges of the tetrahedron  $SEFG$  are perpendicular to the faces of the tetrahedron?



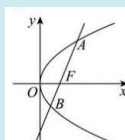
**Visual Aids:**

**Caption of Visual Aids:** Make a four prism S-EFG and unfold it with SG GE GF to form the original square  $SG_1G_2G_3$

Table 14: Example of mathematical problem with solid geometry as visual aid.

**ID:** 394

**Question:** Through the parabola  $C : y^2 = 2px (p > 0)$  The focus of  $F$ , as the slope of  $2\sqrt{2}$  line  $l$  and parabola  $C$  intersect at two points  $A, B$ , If  $\vec{AF} = \lambda \vec{FB}$ ,  $\lambda = ( \quad )$



**Visual Aids:**

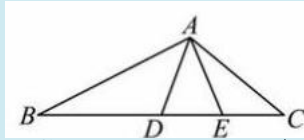
**Caption of Visual Aids:** Let me draw the parabola  $C : y^2 = 2px (p > 0)$ , focus  $F$ , as the slope of  $2\sqrt{2}$  line  $l$  and parabola  $C$  intersect at two points  $A, B$

Table 15: Example of mathematical problem with function graph as visual aid.

**ID:** 30

**Data Source:** Chinese Middle School Practice

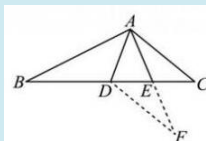
**Visual Context:**



**Caption of Visual Context:** Figure: In  $\triangle ABC$ ,  $BD = DC = AC$ ,  $AE$  is the center line of  $\triangle ADC$ .

**Question:** It is known that, as shown in the figure:  $\triangle ABC$ ,  $BD = DC = AC$ ,  $AE$  is the center line of  $\triangle ADC$ : judge if:  $AB = 3AE$ .

**Visual Aids:**



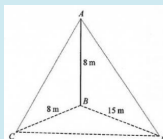
**Caption of Visual Aids:** Extend  $AE$  to  $F$ , connect  $EF = AE$  to  $DF$

Table 16: Example of mathematical problem classified into “Easy” category

**ID:** 650

**Data Source:** HKDSE Examination

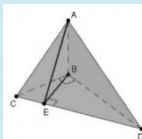
**Visual Context:**



**Caption of Visual Context:**  $AB$  is a vertical pole standing on the horizontal ground  $BCD$ , where  $\angle CBD = 90^\circ$

**Question:** In the figure,  $AB$  is a vertical pole standing on the horizontal ground  $BCD$ , where  $\angle CBD = 90^\circ$ . If the angle between the plane  $ACD$  and the horizontal ground is  $\theta$ , then  $\tan \theta =$

**Visual Aids:**



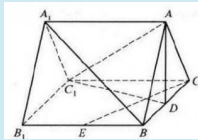
**Caption of Visual Aids:** From point  $B$  to point  $CD$ , the perpendicular intersects point  $CD$  at point  $E$ , connecting  $AE$ ,  $BE$

Table 17: Example of mathematical problem classified into “Medium” category

**ID:** 1024

**Data Source:** Chinese Mathematical Olympiad (CMO)

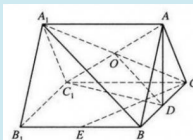
**Visual Context:**



**Caption of Visual Context:** In the right triangular prism  $ABC - A_1B_1C_1$ ,  $AB = AC = 5$ ,  $D, E$  are the midpoints of  $BC, BB_1$  respectively, and the quadrilateral  $B_1BCC_1$  is a square with side length 6.

**Question:** As shown in the figure, in the rectangular triangular prism  $ABC - A_1B_1C_1$ ,  $AB = AC = 5$ ,  $D, E$  are the midpoints of  $BC, BB_1$  and the quadrilateral  $B_1BCC_1$  is a square of side 6. determine that:  $A_1B //$  is in the plane of  $AC_1D$ .

**Visual Aids:**



**Caption of Visual Aids:** Connect  $A_1C$ , intersect with  $AC_1$  at  $O$  point, connect  $OD$

Table 18: Example of mathematical problem classified into “Hard” category

Category	Detail
Auxiliary Line	<p>1.1 Connect the points a,b to make the line segment</p> <p>1.2 Connect points a,b and extend them to intersect the CD line at point E.</p> <p>1.3 Make a vertical line AB through point a.</p>
Rectangular Three-Dimensional Coordinate System	<p>2.1 Establish a three-dimensional rectangular coordinate system with point o as the origin and oa as the x-axis positive direction and ob as the y-axis and oc as the z-axis positive direction.</p>
Rectangular Coordinate System	<p>3.1 Take point o as the origin oa as the x-axis positive direction ob as the y-axis Establish a right-angle coordinate system</p> <p>3.2 With point o as origin oa as x-axis positive direction y-axis positive direction as x-axis rotated 90 degrees counterclockwise/clockwise</p>
Function Graph	<p>4.1 Draw the <math>y^2 = ax</math> image in the real coordinate system with 0 as the origin.</p> <p>4.2 Draw the <math>y^2 = ax</math> image in the real coordinate system with 0 as the origin and the <math>y=ax</math> image intersecting at point k in the first quadrant and at point D in the second quadrant.</p>
Riemann integral problem	<p>5.1 Draw the graph of [function] function in interval [interval] and draw the Riemann integral graph in units of [unit] with values on the right/left/middle side.</p>

Table 19: Caption writing templates of various visual aid types for annotators' reference.

## F Detail Experiment Settings

### F.1 Hyperparameters

We utilize the default inference settings for each LLMs and LMMs in our experiments. Only specific hyperparameters that are necessary to clarify are listed in Table 23 and 24. We conduct open source model inference based on (Wolf et al., 2020). We report a single run score for each experiment.

### F.2 Reasoning Prompter

We list the ICL prompts for assigning different models to perform reasoning under six task settings in mathematical domain: 1) CQ2A 2) CQ2VA 3) CQpV2A 4) pCQ2A 5) pCQ2VA 6) pCQpV2A. Task instructions for each task are listed in Table 25. ICL examples can be found at our code open sourced after acceptance.

### F.3 Instruction across Answer Types

To facilitate accuracy evaluation, each sample is bound with with non-ambiguous result across integer, fraction, decimal and choice. We define choice as answer type for multiple choice and true/false problems. The specific task instruction for each output answer type is shown in Table 26.

### F.4 Detail of Evaluation Dimensions

These dimensions are defined as follows:

- **Logical Rigor (LogiR):** Assesses the coherence and completeness of the logical chain, ensuring each step strictly follows from previous ones without logical leaps.
- **Tool-Free Feasibility (ToolF):** Determines if all calculations can be performed via mental arithmetic or standard pen-and-paper methods, without relying on external computing tools.
- **Information Traceability (InfoT):** Verifies that all information used originates from the problem statement or common knowledge, preventing the use of external, unstated data (i.e., data leakage).
- **Method Generality (GeneR):** Evaluates whether the solution method is a general framework applicable to similar problems, rather than a memorized answer specific to the given instance.

### F.5 Visual Aid Extraction

Reasoning prompter instruct models to provide visual aids, reasoning and answer following certain format. We can directly extract visual aids and followup reasoning from output generated by LLM and LMMs with sufficient zero-shot reasoning capability (eg. Doubao-Seed-1.6, GPT-4.1). However, few weaker models (eg. llama2, LLaVA-Next-Mistral-7B) fail to provide standardized output. We regard these output as both visual-aid and followup reasoning for later answer extraction and similarity evaluation respectively.

### F.6 Answer Extraction Prompter

Answer extraction is conducted upon the extracted reasoning text from model output. We provide five ICL examples instructing GPT-4o mini to extract answer from varied model outputs. Examples are shown in Table 27.

## G Robustness Analysis of LLM-as-a-Judge Evaluation

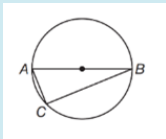
This appendix provides additional evidence for the robustness of the process evaluator used in PVA and SPRS. Since our framework relies on LLM-as-a-judge rather than human annotation, we assess evaluator reliability through repeated-scoring stability under sampling perturbations and consistency across different judge models.

### G.1 Experimental Setup

For robustness evaluation, we randomly sample 100 solution processes from the benchmark and repeatedly score the same responses under different decoding settings. We vary temperature in  $\{0.4, 0.6, 0.8\}$  and top- $p$  in  $\{0.4, 0.7, 1.0\}$  while keeping the evaluation prompt fixed. Each response is scored on the same 0-10 scale for the four process dimensions used in the main paper: Logical Rigor, Tool-Free Feasibility, Information Traceability, and Generality. GPT-4.1 serves as the primary judge in the main experiments, while Qwen3-Max is used as an auxiliary judge for robustness comparison.

For repeated-scoring analysis, we report the mean  $\mu$ , standard deviation  $\sigma$ , coefficient of variation ( $CV = \sigma/\mu$ ), and Spearman rank correlation  $\rho$ . The Spearman correlation measures the agreement between a single review vector and the mean score vector obtained from repeated reviews of the

**Visual Context:**

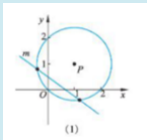


**Question:**  $\overline{AB}$  is a diameter,  $AC = 8$  inches, and  $BC = 15$  inches. Find the radius of the circle.

**GPT4 Reasoning:** To solve for the radius of the circle, we can use the fact that triangle ACB is a right triangle because AB is the diameter of the circle (Thales' theorem). Given: •  $AC=8$  inches •  $BC=15$  inches Using the Pythagorean theorem in triangle ACB:  $AB^2=AC^2+BC^2$  Substitute the given values:  $AB^2=8^2+15^2=64+225=289$  Thus,  $AB=\sqrt{289}=17$  Since AB is the diameter, the radius  $r$  is half of AB:  $r=17/2=8.5$  The correct answer is (C) 8.5.

Table 20: An MPS example with visual context from MathVista and reasoning provided by GPT4V.

**Visual Context:**



**Question:** It is known that the line  $m : 3x + 4y - 2 = 0$  and the circle  $P : x^2 - y^2 - 2x - 2y = 0$ . (2) From the graph drawn in (1), determine the relationship between the line  $m$  and the circle  $P$ . If they intersect, find the length of the chord of the line  $m$  intercepted by the circle  $P$ . If they are tangent or disjoint, give a proof.

Table 21: An example of using only arithmetic methods to solve mathematical problems

**GPT4 Reasoning:** Step 1: Solve  $m$  for  $y$  to find the slope and intercept: ...  $3x + 4y - 2 = 0 \Rightarrow y = -\frac{3}{4}x + \frac{1}{2}$ .

Step 2: Complete the square for circle  $P$  to find its center and radius:  $x^2 - 2x + y^2 - 2y = 0 \Rightarrow (x - 1)^2 + (y - 1)^2 = 2$ .

Step 3: Substitute  $y = -\frac{3}{4}x + \frac{1}{2}$  into the circle's equation and solve for  $x$ .

\*\*Step 4:\*\* If two points of intersection exist,  $A(x_1, y_1)$  and  $B(x_2, y_2)$ , use the distance formula to find the chord length:

$$\text{Chord Length} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

The final answer will be the length of the chord, provided as an integer in LaTeX format. Let's carry out the calculations to find the result ...

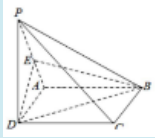
same solution set. For cross-judge analysis, we report Pearson correlation and mean shift  $\Delta\mu$ , where a negative  $\Delta\mu$  indicates that Qwen3-Max assigns lower scores on average than GPT-4.1.

PVA is a binary pass/fail indicator derived from the judge's holistic process-verification decision. In repeated-scoring analysis on a fixed sample set, we primarily report its mean pass rate as a descriptive reference. Variance-based statistics and rank-based correlations are less informative for this binary outcome and are therefore omitted.

## G.2 Repeated-Scoring Stability

Tables 28 and 29 summarize the repeated-scoring stability of GPT-4.1 and Qwen3-Max, respectively. Overall, GPT-4.1 shows low relative variation across the four evaluation dimensions, with coefficients of variation generally in the range of 0.10-0.20, together with high rank-order consistency ( $\rho = 0.84-0.94$ ). The aggregated SPRS also remains stable, achieving a Spearman correlation of 0.94 despite moderate absolute variance introduced by multiplicative aggregation. These results indicate that the proposed process evaluation is not overly sensitive to moderate sampling perturbations

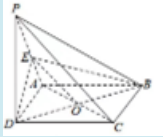
**Visual Context:**



**Question:** Given that  $PD \perp \text{plane } ABCD$ ,  $ABCD$  is a square, and  $E$  is the midpoint of  $PA$ , judge : (1)  $PC \parallel \text{plane } EBD$

Table 22: A successful direct visual-aided reasoning example

**GPT4 Constructed Visual Aids:** Connect  $PC$  and  $AC$ , which intersect with  $BD$  at point  $O$ . Draw a line  $EO$  from point  $E$  to  $O$ .



**GPT4 Reasoning:** Since  $E$  is the midpoint of  $PA$  and  $O$  is the midpoint of diagonal  $AC$  of square  $ABCD$ ;  $\therefore EO$  is parallel to  $PC$  (as midlines in a trapezoid);  $\therefore PC \not\subset \text{plane } EBD$  and  $EO \subset \text{plane } EBD$ ;  $\therefore PC$  is parallel to plane  $\text{plane } EBD$ . So the answer is True.

Model	Hyperparameters
GPT3.5	model = gpt-3.5-turbo, temperature = 0, max_tokens = 4000
GPT-4	model = gpt-4-turbo-2024-04-09, temperature = 0, max_tokens = 4000
Qwen-VL-Plus	model = qwen-vl-plus, temperature = 0.7, max_tokens = 4000
Gemini-Pro-Vision	model = gemini-pro-vision, temperature = 0.4, max_tokens = 4000
Claude-3-Sonnet	model = claude-3-sonnet, temperature = 1, max_tokens = 4000
GPT4V	model = gpt-4-vision-preview, temperature = 0, max_tokens = 4000
GPT4.1	model = gpt-4-1-2025-04-14, temperature = 0, max_tokens = 4000
o4-mini	model = o4-mini-2025-04-16, temperature = 0, max_tokens = 4000
Doubao-Seed-1.6	model = doubao-seed-1-6-250615, temperature = 0, max_tokens = 4000

Table 23: Hyperparameter for close LLMs and LMMs in experiments.

and preserves reliable relative rankings among solution processes.

Qwen3-Max exhibits somewhat larger variability than GPT-4.1, especially on Information Traceability and Generality, but still maintains acceptable rank-order consistency across repeated reviews. This suggests that evaluator stability may

vary across judge models, while the overall ordinal structure of process-quality assessment remains reasonably robust.

### G.3 Cross-Judge Consistency

To further assess whether the observed evaluation trends are specific to a single judge, we compare

Model	Hyperparameters
Llama2-7B	model = Llama-2-7b-chat-hf, precision = bfloat16, temperature = 1.0, max_tokens = 4000
Mistral-7b-Instruct-v0.2	model = mistral-7b-instruct-v0.2, precision = bfloat16, temperature = 1.0, max_tokens = 4000
LLaVA-Next-Mistral-7B	model = llava-v1.6-mistral-7b-hf, precision = float16, temperature = 1.0, max_tokens = 4000
InternLM-XComposer2-VL	model = internlm-xcomposer2-v1-7b, precision = float16, temperature = 1.0, max_tokens = 4000
VL-Cogito	model = VL-Cogito, precision = float16, temperature = 0, max_tokens = 4000
Qwen2.5-VL	model = Qwen2.5-VL-72B-Instruct, precision = float16, temperature = 0, max_tokens = 4000
InternVL3	model = InternVL3-78B, precision = float16, temperature = 0, max_tokens = 4000

Table 24: Hyperparameter for open LLMs and LMMs in experiments.

GPT-4.1 and Qwen3-Max on the same set of solution processes. Table 30 reports Pearson correlations and mean shifts between the two judges.

The two evaluators show strong agreement in relative scoring patterns, with Pearson correlations of 0.72-0.80 across the four process dimensions and 0.80 for SPRS. At the same time, the mean shifts are consistently negative, indicating that Qwen3-Max tends to assign lower absolute scores than GPT-4.1. This pattern suggests that the main cross-model difference lies in scoring scale calibration rather than disagreement about comparative solution quality. Therefore, while absolute score magnitudes may vary across judges, the comparative trends and ranking structure underlying our analysis remain substantially consistent.

Overall, the robustness analyses in this appendix support the use of GPT-4.1 as a stable judge for process-level evaluation in VisAidMath. The evaluator remains reasonably stable under sampling perturbations, and the main comparative trends are preserved across different judge models. These findings reduce the likelihood that the performance gaps from ACCU to PVA/SPRS reported in the main paper are artifacts of evaluator randomness or judge-specific scoring bias.

## H More Experimental Results

More experimental results are in Figure 9.

### H.1 Results of other tasks

#### H.2 Quantitative Analysis

As shown in Figures 10, 11, and 12, the accuracy scores of LMMs and LLMs across math branches show that GPT4V and Gemini-Pro-Vision excel in different tasks with visual aids.

##### H.2.1 Performance across Math Branches

Figure 10, 11, and 12 illustrate the accuracy scores of LMMs upon three tasks with image input across four math branches. GPT4V outperforms other models in problems within “plane geometry”, “analytic geometry”, and “calculus and functions” branches. Gemini-Pro-Vision achieves the highest score on solid geometry problems. Notably Claude-3-Sonnet and InternLM-XComposer2-VL both achieves comparable results toward GPT4V in “solid geometry” branch when reason with provided visual-aided, exhibiting robustness and enhanced capabilities in spatial understanding and visual reasoning under “solid geometry”. GPT4V underperforms in direct visual-aided reasoning, exhibiting significant deficiency processing implicit visual information. Gemini-Pro-Vision performs better at “analytic geometry” and “calculus and functions” with provided visual aids, demonstrating better understanding of visual context within these mathematical branches.

The accuracy scores of both LLMs and LMMs

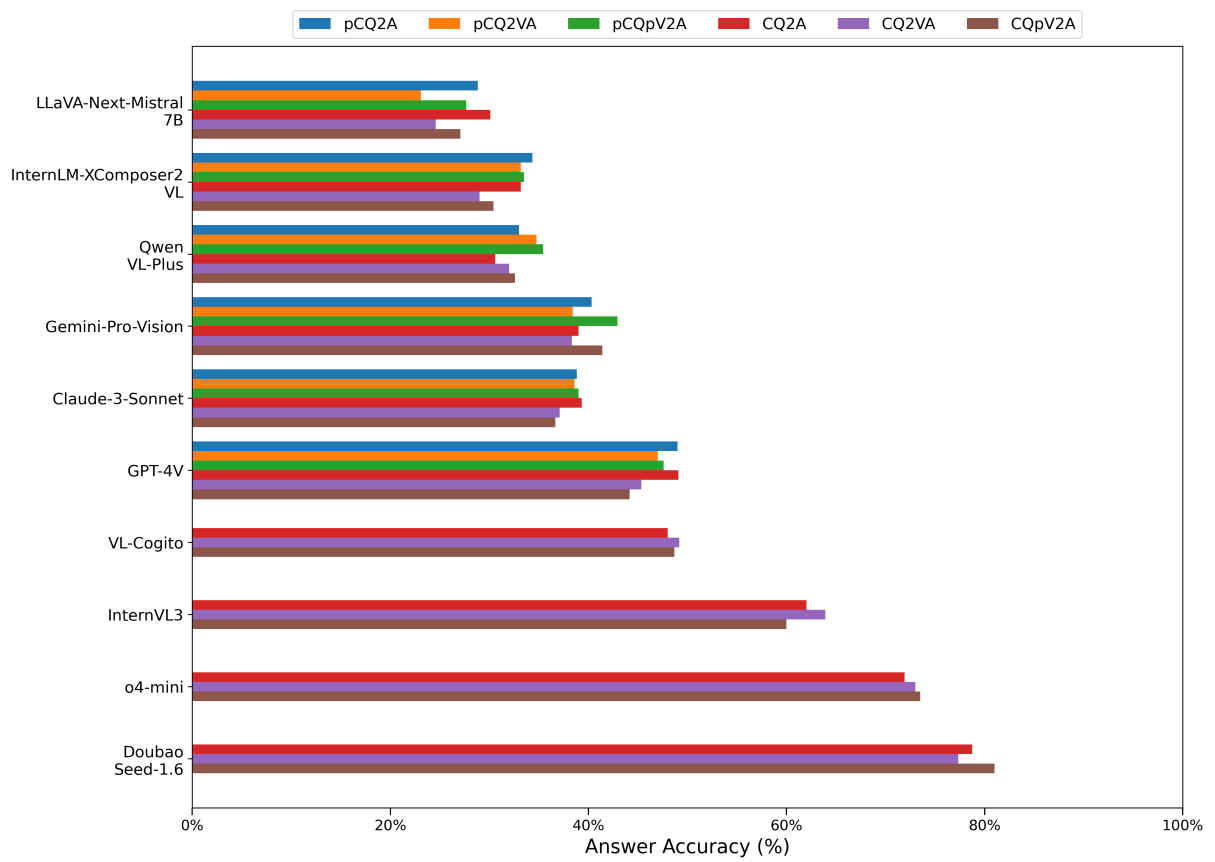


Figure 9: CQ2A is corresponding to GR, CQ2VA is corresponding to D-VAR, and CQpV2A is corresponding to I-VAR.

Answer Type	With Visual Context
CQ2A	Please read the following math problem and the related image. After that,
CQ2VA	Please read the following math problem and the related image, then conceive the additional mathematical diagram or visual aid upon provided image to help illustrate the problem, clarify the problem or assist in finding solution. The drawing shape includes auxiliary line, plane geometry graph, solid geometry graph, function graph, plane coordinate and three-dimensional coordinate. These additional drawings can enhance understanding of the problem and potentially find new insights or approaches to solving it. Write drawing description of these additional diagram in text, and express math formula with tex gramma. After that, reason based on the drawing description,
CQpV2A	Please read the following math problem and the related image. Read the provided drawing description of additional mathematical diagram or visual aid upon provided image in latex format, which helps illustrate the problem, clarify the problem or assist in finding solution. The drawing shape includes auxiliary line, plane geometry graph, solid geometry graph, function graph, plane coordinate and three-dimensional coordinate. These additional drawings can enhance understanding of the problem and potentially find new insights or approaches to solving it. After that, based on the drawing description,
pCQ2A	Please read the following math problem and captions of related visual context. After that,
pCQ2VA	Please read the following math problem and captions of related visual context, then conceive the additional mathematical diagram or visual aid upon provided image to help illustrate the problem, clarify the problem or assist in finding solution. The drawing shape includes auxiliary line, plane geometry graph, solid geometry graph, function graph, plane coordinate and three-dimensional coordinate. These additional drawings can enhance understanding of the problem and potentially find new insights or approaches to solving it. Write drawing description of these additional diagram in text, and express math formula with tex gramma. After that, reason based on the drawing description,
pCQpV2A	Please read the following math problem and captions of related visual context. Read the provided drawing description of additional mathematical diagram or visual aid upon provided image in latex format, which helps illustrate the problem, clarify the problem or assist in finding solution. The drawing shape includes auxiliary line, plane geometry graph, solid geometry graph, function graph, plane coordinate and three-dimensional coordinate. These additional drawings can enhance understanding of the problem and potentially find new insights or approaches to solving it. After that, based on the drawing description,

Table 25: Instructions for different mathematical problem solving tasks with visual context inside input.

on three tasks with image caption for visual context across math branches is shown in Figure 13, 14, and 15. In text-only reasoning, GPT4 outperforms other models across most branches and tasks. GPT4V and Gemini-Pro-Vision achieve the highest

score at “plane geometry” and “solid geometry” respectively in pCQ2A task. Gemini-Pro-Vision also attains the best score at “solid geometry” branch, manifesting robust visual context utilization within “solid geometry” branch. GPT4V and GPT4 share

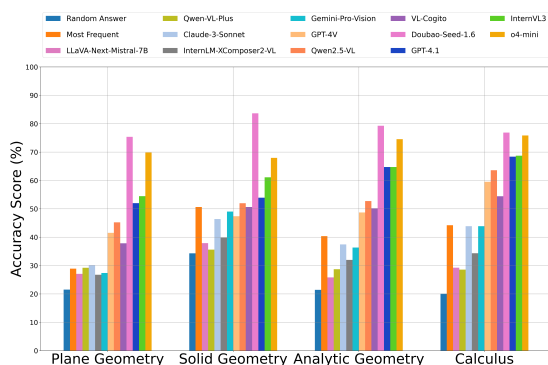


Figure 10: Accuracies of LLMs and LMMs upon CQ2A (General Text-Only Reasoning) task across math branches.

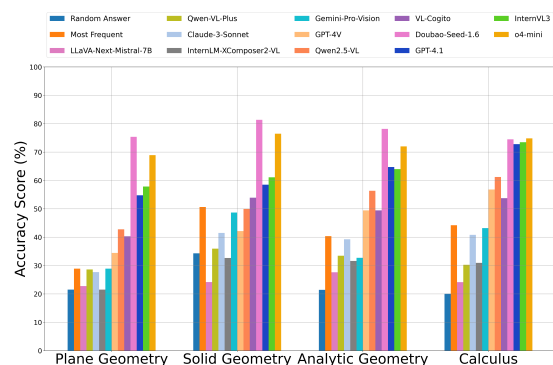


Figure 11: Accuracies of LLMs and LMMs upon CQ2VA (Direct Visual-Aided Reasoning) task across math branches.

Answer Type	Task Instruction
Integer	Answer the question requiring a integer answer in latex format and provide the final value at the end (e.g., -1, 0, 1, 2, 3).
Decimal	Aswer the question requiring a decimal answer in latex format and provide the final value at the end. Make sure the decimal answer is a floating-point number with three decimal place (e.g., 1.234, 2.345, 3.456).
Fraction	Answer the question requiring an fraction answer in latex format and provide the final value at the end. Make sure the fraction answer use '/' as fraction bar and both numerator and denomi-nator are integers (e.g., 1/2, 2/3, 3/4).
Choice (Multiple-Choice)	Answer the question in latex format and provide the correct option at the end. Make sure the option answer can only be single capital letter (e.g., A, B, C, D).
Choice (True/False)	Answer the question in latex format and provide the correct option at the end. Make sure the option answer can only be word 'True' or 'False'.

Table 26: Instructions for mathematical problem solving toward different answer types.

similar performances over “plane geometry” problems. However, performance of these two models deviates when conduct indirect reasoning. However, GPT-4V achieves significantly lower accuracy on “analytic geometry” branch, demonstrating difficulty for GPT-4V to handle additoinal visual information on “analytic geometry” branch. Gemini-Pro-Vision outperform other models after introduce visual aids into reasoning at “solid geometry” category, indicating better three-dimensional spatial information understanding and processing. Overall, more models perform reasoning better with visual aids on “solid geometry” and “analytic ge-

ometry” problems, possibly because visual aids in these problems are bound tighter to the reasoning path. In opposite, visual aids within “plane geometry” problems still often leave broad decision space, thus making it harder to utilize the additonnal visual context. Since various problems of “calculus and functions” can often be solved by generic method, most models maintain the highest accuracy in this branch.

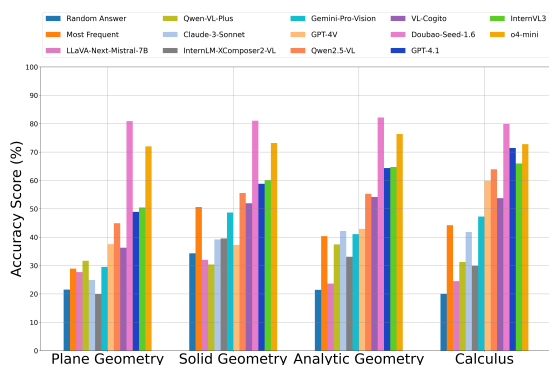


Figure 12: Accuracies of LLMs and LMMs upon CQpV2A (Indirect Reasoning) task across math branches.

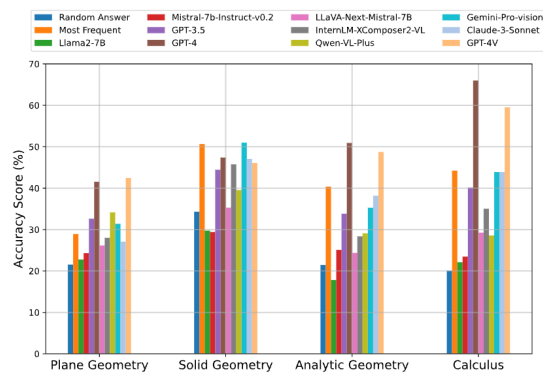


Figure 13: Accuracies of LLMs and LMMs upon pCQ2A (General Text-Only Reasoning) task across math branches.

Element	Prompt
Example 1	<p><b>Instruction:</b> Answer the question requiring a integer answer and provide the final value at the end (e.g., -1, 0, 1, 2, 3).</p> <p><b>Question:</b> Acute Angle <math>\triangle ABC</math>, <math>BE</math> vertical <math>AC</math> at <math>E</math>, <math>CD</math> vertical <math>AB</math> at <math>D</math>, given acute Angle <math>\triangle ABC</math>, <math>BE</math> vertical <math>AC</math> at <math>E</math>, <math>BCD</math> vertical <math>A</math> in <math>D</math>, <math>BC = 25</math>, <math>CE = 7</math>, <math>BD = 15</math>, <math>BE</math>, <math>CD</math> to <math>H</math>, connection <math>DE</math>, <math>DE</math> for circle diameter, and <math>AC</math> to another <math>F</math>, Then the length of <math>AF</math> is</p> <p><b>Model response:</b> Connect <math>DF</math>, then <math>\angle DFE = \angle BEC = 90^\circ</math>, so <math>DF \parallel BE</math>. In Rt <math>\triangle BEC</math>, <math>BC = 25</math>, <math>CE = 7</math>, so <math>BE = 24</math>, <math>\cos \angle ACB = \frac{7}{25}</math>, <math>\sin \angle ACB = \frac{24}{25}</math>, in Rt <math>\triangle BCD</math>, <math>BC = 25</math>, <math>BD = 15</math>, so <math>\cos \angle ABC = \frac{3}{5}</math>, <math>\sin \angle ABC = \frac{4}{5}</math>. So <math>\sin A = \sin(\angle ABC + \angle ACB) = \sin \angle ABC \cos \angle ACB + \cos \angle ABC \sin \angle ACB = \frac{4}{5}</math>, in Rt <math>\triangle ABE</math>, <math>\sin A = \frac{BE}{AB} = \frac{24}{AB} = \frac{4}{5}</math>, we get <math>AB = 30</math>, and <math>BD = 15</math>, so the point <math>D</math> is the midpoint of <math>AB</math>, and <math>DF</math> is the median of <math>\triangle ABE</math>, so <math>AF = \frac{1}{2}AE = \frac{1}{2}\sqrt{30^2 - 24^2} = 9</math>. So the final answer is 9</p> <p><b>Extracted answer: 9</b></p>
Example 2	<p><b>Instruction:</b> Answer the question requiring a decimal answer and provide the final value at the end. Make sure the decimal answer is a floating-point number with three decimal place (e.g., 1.234, 2.345, 3.456)</p> <p><b>Question:</b> In a right angle <math>\triangle ABC</math>, <math>\angle DAC = 45^\circ</math>, <math>AD = 13</math>, <math>AB = 12</math>. Calculate the length of <math>CD</math>.</p> <p><b>Model response:</b> Let <math>DE = x</math>, then <math>CE = AE = 13 + x</math>, <math>\frac{12}{13+x} = \frac{5}{x}</math> <math>x = \frac{65}{7}</math>, <math>DE = \frac{65}{7}</math> <math>CE = \frac{156}{7}</math>, <math>CD = \sqrt{DE^2 + CE^2} = \frac{169}{7} = 24.1</math></p> <p><b>Extracted answer: 24.143</b></p>
Example 3	<p><b>Instruction:</b> Answer the question requiring an fraction answer and provide the final value at the end. Make sure the fraction answer use '/' as fraction bar and both numerator and denominator are integers (e.g., 1/2, 2/3, 3/4).</p> <p><b>Question:</b> Linear <math>x + y = \sqrt{3}a</math> and round <math>x^2 + y^2 = a^2 + (a - 1)^2</math> intersect at point of <math>a, B</math>, <math>O</math> is the origin of coordinates, if <math>\triangle AOB</math> is triangle, the real value of <math>a</math></p> <p><b>Model response:</b> From the question, the radius of the circle <math>r = \sqrt{a^2 + (a - 1)^2}</math>; the distance from the centre of the circle to the line <math>x + y - \sqrt{3}a = 0</math> <math>d = \frac{ \sqrt{3}a }{\sqrt{2}}</math>; because <math>\triangle AOB</math> is a positive triangle, so <math>d = \frac{\sqrt{3}}{2}r</math>, i.e. <math>\frac{ \sqrt{3}a }{\sqrt{2}} = \frac{\sqrt{3}}{2}\sqrt{a^2 + (a - 1)^2}</math>, which solves <math>a = \frac{1}{2}</math></p> <p><b>Extracted answer: 1/2</b></p>

Table 27: Three ICL examples leveraged to prompt GPT-4 for answer extraction.

## H.2.2 Performance across Complexities

## H.2.3 Performance across Visual Aids

With Image as input across different visual aids required to generate, the accuracy scores of main-stream LMMs under three tasks are listed in Figure 22, 23, and 24. GPT-4V outperforms other models on problems with visual aids “auxiliary line” and “function graph” across all three tasks. In text-only reasoning task, GPT-4V achieves leading performance in text-only reasoning except for problems with rectangular three-dimensional coordinate system as visual aids. After introduce visual aids, Gemini-Pro-Vision significantly outperform other models on problems with solid geometry graph and

three-dimensional rectangular coordinate system as visual aids. Gemini-Pro-Vision also achieves comparable result to GPT-4 in problem solving provided with auxiliary line. Comparing generated and provided visual-aids, overall average accuracy are enhanced saliently on “auxiliary line”, “plane coordinate system”, and “function graph”, exhibiting higher sensitivity in reasoning towards these visual aids.

Figure 25, 26, and 27 demonstrate accuracy scores of LLMs and LMMs on three tasks with image caption as input across visual aids. GPT-4 achieves outstanding scores compared to other models. With golden visual aids provided, GPT-4V attains higher accuracy on problem with “solid

Metric	$\mu$	$\sigma$	CV	Spearman $\rho$
Logical Rigor	8.0844	1.6500	0.2041	0.9408
Tool-Free Feasibility	9.3889	0.9677	0.1031	0.8709
Information Traceability	9.2556	1.5957	0.1724	0.8473
Generality	9.2922	1.3703	0.1475	0.8748
SPRS	0.7130	0.2815	0.3949	0.9387
PVA	0.6911	-	-	-

Table 28: Repeated-scoring stability statistics for GPT-4.1 under different decoding settings.

Metric	$\mu$	$\sigma$	CV	Spearman $\rho$
Logical Rigor	7.8167	2.0910	0.2675	0.8356
Tool-Free Feasibility	8.8967	1.7017	0.1913	0.8523
Information Traceability	8.4956	2.5253	0.2972	0.7644
Generality	8.1456	2.1635	0.2656	0.7543
SPRS	0.5792	0.3465	0.5982	0.8324
PVA	0.6000	-	-	-

Table 29: Repeated-scoring stability statistics for Qwen3-Max under different decoding settings.

Metric	Pearson $r$	$\Delta\mu$
Logical Rigor	0.8024	-0.2678
Tool-Free Feasibility	0.7477	-0.4922
Information Traceability	0.7174	-0.7600
Generality	0.7650	-1.1467
SPRS	0.7986	-0.1338
PVA	-	-0.0911

Table 30: Cross-judge consistency between GPT-4.1 and Qwen3-Max on the same solution set. Negative  $\Delta\mu$  indicates that Qwen3-Max assigns lower scores on average than GPT-4.1.

geometry graph” as visual aids, and Gemini-Pro-Vision well-perform on reasoning with aided “rectangular three-dimensional coordinate system”. Overall scores on “plane geometry graph” exhibit substantial difficulty to employ implicit or explicit information within plane geometry graphs.

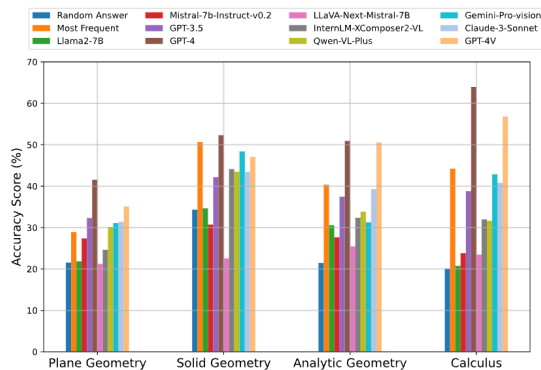


Figure 14: Accuracies of LLMs and LMMs upon pCQ2VA (Direct Visual-Aided Reasoning) task across math branches.

### H.3 N-gram Similarities

We report three n-gram similarities in experiments to fine-grained reveal model capability. For clarity, we first define **hAid** as the generated visual aids from visual-aided generation task. **tAid** is defined

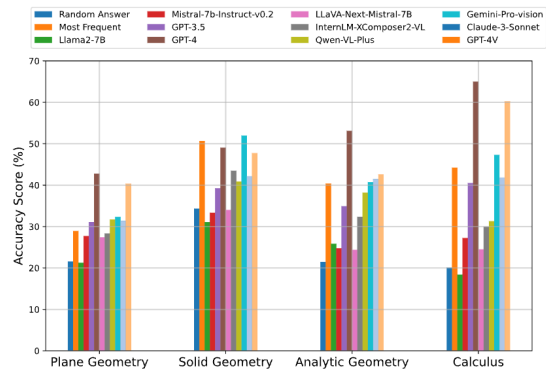


Figure 15: Accuracies of LLMs and LMMs upon pC-QpV2A (Indirect Reasoning) task across math branches.

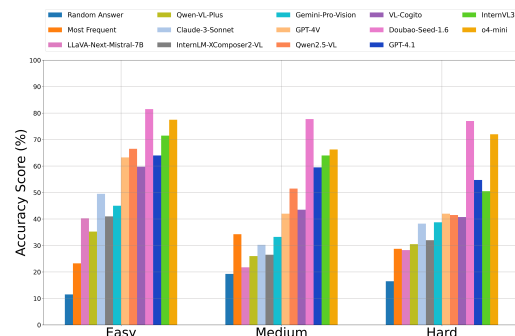


Figure 16: Accuracies of LLMs and LMMs upon CQ2A (General Text-Only Reasoning) task across complexity levels.

Model	ALL	PLG	SDG	AYG	CAL	AXL	RTC	THC	PLG	SDG	FUG
<i>Heuristics Baselines</i>											
Random Answer	24.42	21.54	34.31	21.45	20.07	24.44	20.87	35.16	10.53	32.89	21.50
Frequent Answer	40.83	28.92	50.65	40.36	44.22	32.79	47.25	74.73	20.00	47.73	44.53
<i>Large Language Models (LLMs): Text-Only Input</i>											
Llama2-7B	23.25	22.77	29.74	17.82	22.11	22.80	19.72	28.57	8.42	28.29	21.11
Mistral-7b-Instruct-v0.2	25.58	24.31	29.41	25.09	23.47	24.59	25.46	25.27	6.32	26.32	25.91
GPT3.5	37.75	32.62	44.44	33.82	40.14	37.85	38.30	40.66	17.89	43.42	38.20
GPT4	51.17	41.54	47.39	50.91	65.99	45.45	55.73	59.34	22.11	49.34	61.80
<i>Large Multimodal Models (LMMs): Text-Only Input</i>											
LLaVA-Next-Mistral-7B	28.83	26.15	35.29	24.36	29.25	27.72	28.67	30.77	10.53	35.53	28.79
InternLM-XComposer2-VL	34.33	28.00	45.75	28.36	35.03	32.64	33.49	53.85	13.68	36.18	33.40
Qwen-VL-Plus	33.00	34.15	39.54	29.09	28.57	34.87	30.05	34.07	13.68	43.42	30.52
Gemini-Pro-Vision	40.33	31.38	50.98	35.27	43.88	36.66	41.28	53.85	17.89	49.34	41.84
Claude-3-Sonnet	38.83	27.08	47.06	38.18	43.88	33.38	39.68	57.14	16.84	48.03	42.80
GPT4V	49.00	42.46	46.08	48.73	59.52	43.82	52.75	52.75	22.11	50.00	57.58
<i>Large Multimodal Models (LMMs): Multimodal Input</i>											
LLaVA-Next-Mistral-7B	30.08	27.08	37.91	25.82	29.25	28.32	29.13	41.76	9.47	34.87	28.98
InternLM-XComposer2-VL	33.17	26.77	39.87	32.00	34.35	30.85	32.80	43.96	11.58	36.84	34.93
Qwen-VL-Plus	30.58	29.23	35.62	28.73	28.57	31.15	29.82	27.47	13.68	41.45	30.52
Gemini-Pro-Vision	39.00	27.38	49.02	36.36	43.88	35.32	40.37	52.75	14.74	48.68	42.03
Claude-3-Sonnet	39.33	30.15	46.41	37.45	43.88	34.72	38.99	56.04	16.84	47.37	42.42
GPT4V	49.08	41.54	47.39	48.73	59.52	43.82	53.21	51.65	24.21	51.97	57.97
VL-Cogito	49.17	40.31	53.92	53.74	49.45	45.31	53.85	52.40	55.26	50.23	20.00
Qwen2.5-VL-72B	52.25	42.77	50.00	61.22	56.36	45.01	50.55	62.38	53.95	58.49	23.16
InternVL3.5-38B	63.92	57.85	61.11	73.47	64.00	56.33	72.53	71.21	55.92	67.20	54.74
GPT-4.1	62.42	54.77	58.50	72.79	64.73	56.93	72.53	70.25	56.58	66.51	54.74
o4-mini	73.00	68.92	76.47	74.83	72.00	69.75	87.91	74.09	73.03	71.10	56.84
Doubao-Seed-1.6	77.33	75.38	81.37	74.49	78.18	75.26	90.11	76.97	76.32	75.92	68.42

Table 31: Accuracy scores on **General Reasoning task (GR)** task upon **VisAidMath**. Meanings of all abbreviations are listed below. ALL → overall accuracy. For Mathematical Branch: PLG → plane geometry, SDG → solid geometry, AYG → analytic geometry, CAL: calculus and functions. Visual Aids Types: AXL → auxiliary line, RTC → rectangular coordinate, THC → rectangular three-dimensional coordinate, PLG → plane geometry graph, SDG → solid geometry graph, FUG → function graph. The highest scores in chunk and in general have been emphasized with purple and green to facilitate comparison respectively.

as the golden visual aids. **dReas** demonstrates the direct reasoning output from general reasoning task.

The three similarity scores include 1)  $S_{hAid-dReas}$ : overall n-gram similarity between **hAid** and **dReas**

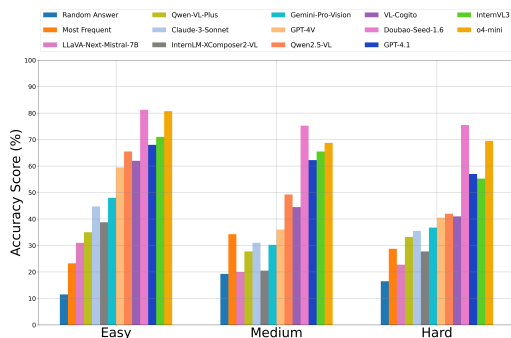


Figure 17: Accuracies of LLMs and LMMs upon CQ2VA (Direct Visual-Aided Reasoning) task across complexity levels.

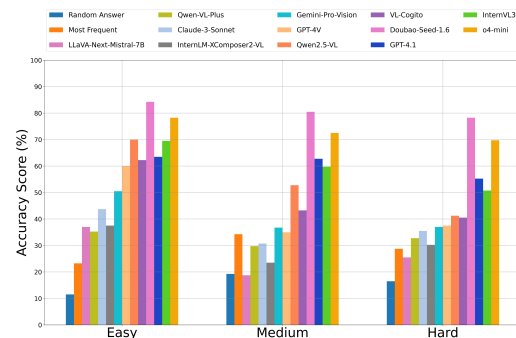


Figure 18: Accuracies of LLMs and LMMs upon CQpV2A (Indirect Reasoning) task across complexity levels.

Model	ALL	PLG	SDG	AYG	CAL	AXL	RTC	THC	PLG	SDG	FUG
<i>Heuristics Baselines</i>											
Random Answer	24.42	21.54	34.31	21.45	20.07	24.44	20.87	35.16	10.53	32.89	21.50
Frequent Answer	40.83	28.92	50.65	40.36	44.22	32.79	47.25	74.73	20.00	47.73	44.53
<i>Large Language Models (LLMs): Text-Only Input</i>											
Llama2-7B	24.08	21.23	31.05	25.82	18.37	25.04	22.71	31.87	7.37	30.26	22.46
Mistral-7b-Instruct-v0.2	28.33	27.69	33.33	24.73	27.21	27.72	27.29	34.07	14.74	32.89	27.26
GPT3.5	36.33	31.08	39.22	34.91	40.48	33.08	37.84	50.55	14.74	39.47	39.73
GPT4	52.17	42.77	49.02	53.09	64.97	46.94	57.11	54.95	20.00	52.63	62.76
<i>Large Multimodal Models (LMMs): Text-Only Input</i>											
LLaVA-Next-Mistral-7B	27.67	27.38	33.99	24.36	24.49	27.42	25.00	29.67	11.58	33.55	25.91
InternLM-XComposer2-VL	33.50	28.31	43.46	32.36	29.93	33.68	32.80	49.45	13.68	41.45	31.86
Qwen-VL-Plus	35.42	31.69	40.85	38.18	31.29	36.51	39.22	40.66	15.79	39.47	34.93
Gemini-Pro-Vision	42.92	32.31	51.96	40.73	47.28	39.79	43.35	57.14	17.89	47.37	45.87
Claude-3-Sonnet	39.00	31.38	42.16	41.45	41.84	35.92	40.14	46.15	17.89	42.11	43.19
GPT4V	47.58	40.31	47.71	42.55	60.20	42.32	47.94	50.55	21.05	55.26	53.93
<i>Large Multimodal Models (LMMs): Multimodal Input</i>											
LLaVA-Next-Mistral-7B	27.08	27.69	32.03	23.64	24.49	27.42	24.31	26.37	11.58	32.89	25.72
InternLM-XComposer2-VL	30.42	20.00	39.54	33.09	29.93	26.97	31.88	40.66	10.53	34.87	32.25
Qwen-VL-Plus	32.58	31.69	30.39	37.45	31.29	33.23	38.99	25.27	16.84	37.50	34.55
Gemini-Pro-Vision	41.42	29.54	48.69	41.09	47.28	37.85	43.81	45.05	14.74	48.03	46.07
Claude-3-Sonnet	36.67	24.92	39.22	42.18	41.84	32.04	40.37	41.76	14.74	43.42	43.76
GPT4V	44.17	37.54	37.25	42.91	59.86	38.60	47.25	36.26	17.89	48.03	53.74
VL-Cogito	48.67	36.31	51.96	53.74	54.18	44.86	47.25	54.89	51.32	50.46	16.84
Qwen2.5-VL-72B	54.67	44.92	55.56	63.95	55.27	47.99	56.04	62.76	54.61	58.03	22.11
InternVL3.5-38B	60.00	50.46	60.13	65.99	64.73	55.44	64.84	66.41	59.87	62.39	34.74
GPT-4.1	60.50	48.92	58.82	71.43	64.36	52.01	75.82	69.48	59.21	68.35	46.32
o4-mini	73.50	72.00	73.20	72.79	76.36	70.34	78.02	75.43	73.03	74.08	70.53
Doubao-Seed-1.6	81.00	80.92	81.05	79.93	82.18	78.84	86.81	81.19	78.95	80.28	75.79

Table 32: Accuracy scores on **Indirect Visual-Aided Reasoning (I-VAR)** task upon **VisAidMath**. Meanings of all abbreviations are listed below. ALL → overall accuracy. For Mathematical Branch: PLG → plane geometry, SDG → solid geometry, AYG → analytic geometry, CAL: calculus and functions. Visual Aids Types: AXL → auxiliary line, RTC → rectangular coordinate, THC → rectangular three-dimensional coordinate, PLG → plane geometry graph, SDG → solid geometry graph, FUG → function graph. The highest scores in chunk and in general have been emphasized with purple and green to facilitate comparison respectively.

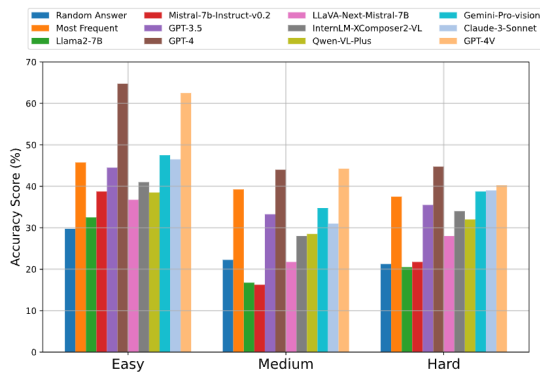


Figure 19: Accuracies of LLMs and LMMs upon pCQ2A (General Text-Only Reasoning) task across complexity levels.

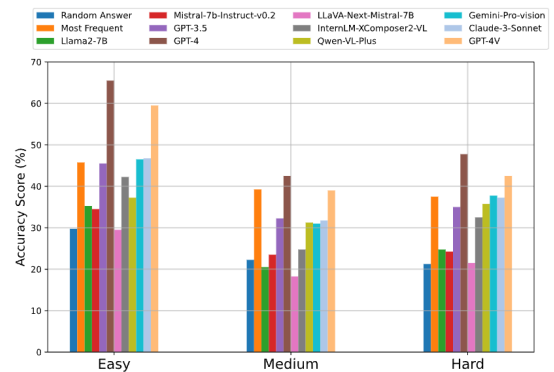


Figure 20: Accuracies of LLMs and LMMs upon pCQ2VA (Direct Visual-Aided Reasoning) task across complexity levels.

2)  $S_{hAid-tAid}$ : target-only n-gram similarity between **hAid** and **tAid** 3)  $S_{dReas-tAid}$  target-only n-gram similarity between **dReas** and **tAid**. To attain these similarity scores, we first need to count each n-gram size for specific text,  $g_n$  represents n-gram chunk:

$$NG = \{Count(g_1), Count(g_2), \dots, Count(g_m)\} \quad (4)$$

$NG_{clip}$  calculates the intersect part of  $NG_i$  and  $NG_j$ , demonstrating explicit n-gram matches.

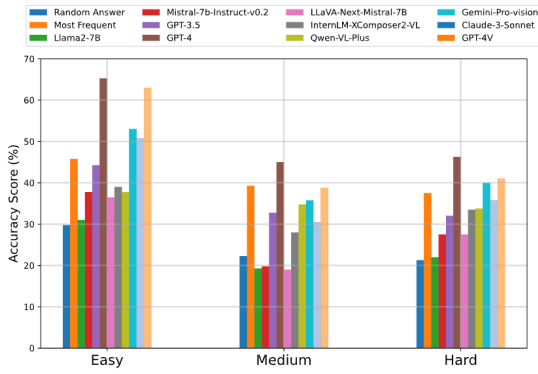


Figure 21: Accuracies of LLMs and LMMs upon pCQpV2A (Indirect Reasoning) task across complexity levels.

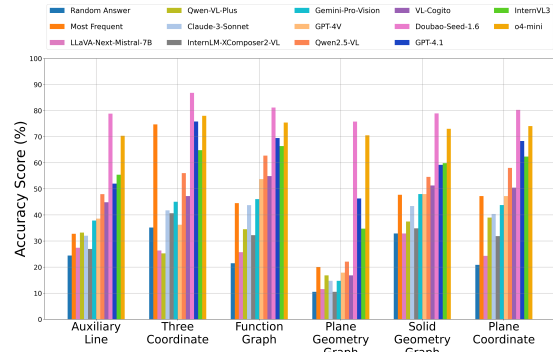


Figure 24: Accuracies of LLMs and LMMs upon CQpV2A (Indirect Reasoning) task across visual aids.

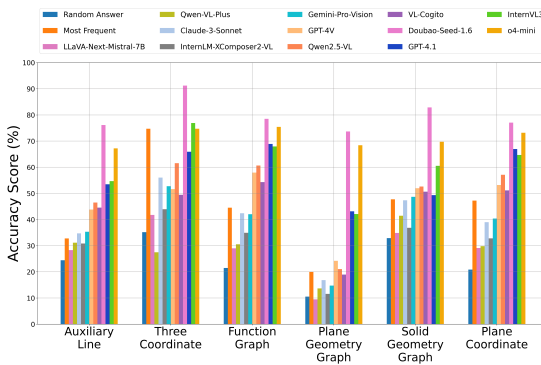


Figure 22: Accuracies of LLMs and LMMs upon CQ2A (General Text-Only Reasoning) task across visual aids.

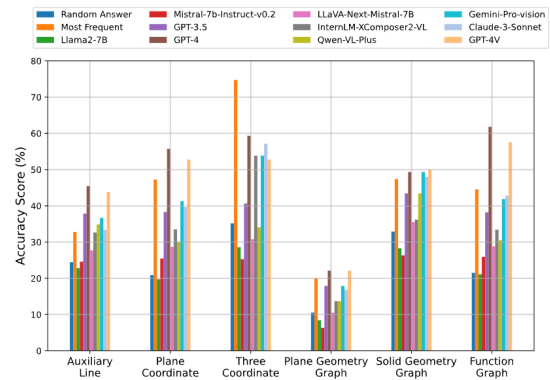


Figure 25: Accuracies of LLMs and LMMs upon pCQ2A (General Text-Only Reasoning) task across visual aids.

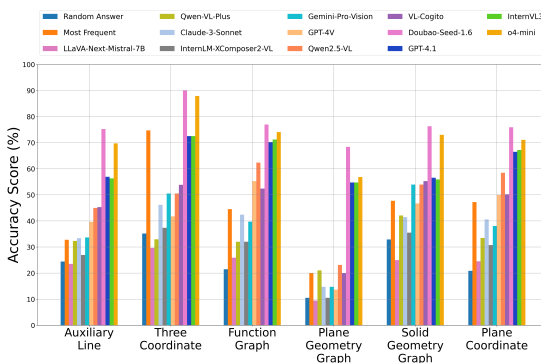


Figure 23: Accuracies of LLMs and LMMs upon CQ2VA (Direct Visual-Aided Reasoning) task across visual aids.

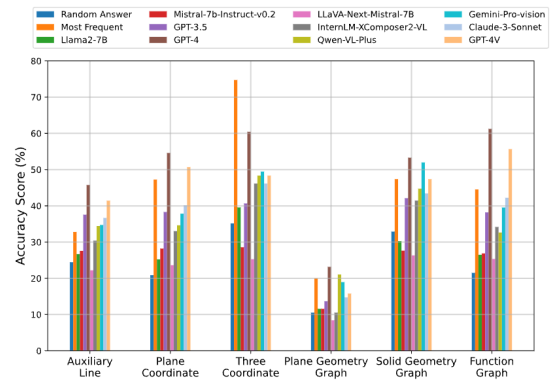


Figure 26: Accuracies of LLMs and LMMs upon pCQ2VA (Direct Visual-Aided Reasoning) task across visual aids.

$NG_{base}$  exhibits essential n-grams to be matched, target-only similarity select  $NG_i$  as  $NG_{base}$  to emphasize match of target n-grams, while overall similarity leverage union of  $NG_i$  and  $NG_j$  as the denominator for similarity calculation.

$$NG_{clip} = NG_i \cap NG_j \quad (5)$$

$$NG_{base} = \begin{cases} NG_i & \text{if target-only simi} \\ NG_i \cup NG_j & \text{if overall simi} \end{cases} \quad (6)$$

Then,

$$[h]S = \frac{|NG_{clip}|}{|NG_{base}|} \quad (7)$$

Qualitative results 4.2 have shown the best performing GPT4 seldom generates visual aids to assist problem solving, which also confirms with quantitative results in Table 33. Since models achieve low  $S_{dReas-tAid}$  and tend to solve mathematical problems without visual aids, it becomes necessary to generate visual aids far from **dReas** in direct visual-aided reasoning task. Model with higher  $S_{hAid-dReas}$  can easily fail in visual aids inference.

## I In-depth Model Analysis

### I.1 General Reasoning Tendency

### I.2 Visual Aid Inference Capability

As shown in Tables 40 and 41, GPT4V demonstrates varying performance in generating visual aids, both correctly and incorrectly.

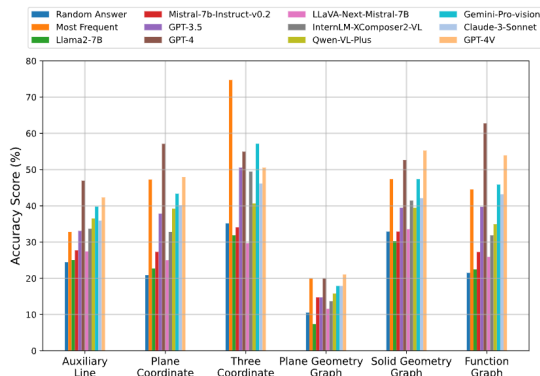


Figure 27: Accuracies of LLMs and LMMs upon pC-QpV2A (Indirect Reasoning) task across visual aids.

Model	$S_{hAid-dReas}$	$S_{hAid-tAid}$	$S_{dReas-tAid}$
<i>Large Language Models (LLMs): Text-Only Input</i>			
Llama2-7B	14.73	5.26	3.04
Mistral-7b-Instruct-v0.2	57.21	5.84	4.88
GPT3.5	10.13	4.36	4.29
GPT4	2.37	4.21	3.47
<i>Large Multimodal Models (LMMs): Text-Only Input</i>			
LLaVA-Next-Mistral-7B	29.59	1.96	4.09
InternLM-XComposer2-VL	76.02	4.88	4.84
Qwen-VL-Plus	11.03	1.89	0.85
GeminiPro-vision	7.35	5.37	3.11
Claude-3-Sonnet	2.37	4.66	2.93
GPT4V	1.52	4.03	3.00
<i>Large Multimodal Models (LMMs): Multimodal Input</i>			
LLaVA-Next-Mistral-7B	37.43	1.99	4.04
InternLM-XComposer2-VL	61.43	4.82	4.73
Qwen-VL-Plus	13.71	1.97	0.95
GeminiPro-vision	6.93	5.48	3.08
Claude-3-Sonnet	2.26	4.61	2.95
GPT4V	1.91	3.98	3.03

Table 33: Three types of similarity scores reported across different models and modalities.

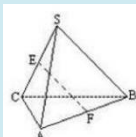
## J Supplementary Quantitative Analysis

As shown in Figure 28a and Figure 6b, we analyze reasoning patterns, hallucination correlation, and failure cases of visual-aid generation.

**Reasoning Pattern** We evaluate only the model outputs with correct answers, examining how GPT4V responds to questions without being instructed to generate visual aids as an intermediate step. We found that GPT4V did not prefer generating visual aids to simplify reasoning complexity. Expect for generating answers due to hallucinations (33.2%), the model searches for problem-solving chains based on visual-irrelevant logic, such as reasoning with pure arithmetic calculations and backward reasoning. An example of solving the problem with pure arithmetic calculation can be found in Table 38.

**Failure Analysis of Visual Aids Generation** We investigated the causes of poorly generated visual aids in the CQ2VA task (i.e., generating visual aids before reasoning). Based on the analysis results, we provide potential research direction for improving the quality of visual aid generation in section A.

**Correlation between Visual-Aid and Hallucination** We analyzed the effectiveness of visual aids in reducing hallucinations. We first categorize the error rates of generated visual aids as low, medium, and high. The hallucination level of the subsequent reasoning is defined as low, medium high. We collect and count combinations of each category and visualize in Figure 28a, including the percentage



ID: 625

**Question:** As shown in the figure, the lateral prongs of the square triangular pheasant  $SABC$  are equal to the lengths of the sides of the base, and if  $E, F$  is the midpoint of  $SC, AB$ , then the angle between  $EF$  and  $SA$  is equal to ( ).

**Choices:** A.  $90^\circ$  B.  $60^\circ$  C.  $45^\circ$  D.  $30^\circ$

Table 34: Example of GPT4V solve mathematical problem by creating visual aids to reveal implicit visual information without instructed.

**Reasoning:** Connect  $SE$  and  $SF$ ,  $\because E, F$  are midpoints of  $SC, AB$  respectively, and given that  $SABC$  is a square triangular pheasant with lateral prongs equal to the base, we know  $SE = SF = \frac{1}{2}SC = \frac{1}{2}AB$ . Since  $AB \perp BC$  and  $SC \perp BC$ ,  $SE \perp SF$ , and triangle  $SEF$  forms a right triangle with  $EF$  as the hypotenuse. Since  $SE = SF$ ,  $\triangle SEF$  is also an isosceles right triangle, so  $\angle SEF = 45^\circ$ . Since  $SA \perp AB$  and  $SE \perp SF$ , plane  $SAE$  is perpendicular to plane  $SEF$ . Therefore, the angle between  $EF$  and  $SA$  is also  $45^\circ$ . So, the final answer is C.  $45^\circ$ .

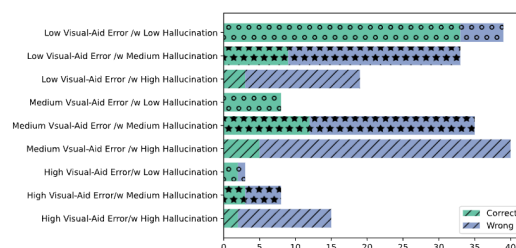
Table 35: Example of GPT4V solve mathematical problem by performing text-only reasoning to achieve correct result.

of correct and incorrect answers. We observe that models with lower visual-aid error rates (i.e., output more complete visual aids) can generate more truthful outputs instead of hallucinations, thus increasing the success rate of problem-solving. This phenomenon indicates that correct visual aids can effectively alleviate hallucinations during reasoning.

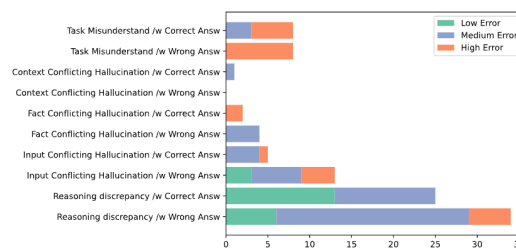
**Failure Analysis of Direct Visual-Aided Reasoning.** We analyzed how models react to poorly generated visual aids. In Figure 6b, the sample size of failure cases is much larger than size of successful cases, highlighting a lack of capability in direct visual-aided reasoning. By comparing sample sizes across different visual aid error rates, we find that poor visual aids with more differences from reference can negatively affect subsequent reasoning. We also summarize the error types in generating visual aids, exploring possible directions to improve LMMs.

## K Related Work

**Benchmark** Numerous benchmarks have been developed to evaluate mathematical reasoning abilities in both textual and multi-modal tasks. These benchmarks primarily rely on textual inference as the reasoning tool. Regarding the text-only task, arithmetic problems with pure numerical expressions (Yuan et al., 2023) and MPS (Yue et al., 2023)



(a) Correlation between visual aid and reasoning hallucination.



(b) Correlation between error reasons of visual aid and answer correctness.

Figure 28: Error analysis of direct visual-aided reasoning task (CQ2VA task, GPT-4V).

have been extensively explored. On the multi-modal side, (Chen et al., 2022; Lu et al., 2021) focus on geometry problems to foster research on spatial understanding and properties deduction. Other multi-modal benchmarks concentrate on general visual contexts, such as bar charts (Masry et al., 2022). More recently, (Lu et al., 2023) established a comprehensive benchmark that incorporates different visual contexts. However, these benchmarks primarily rely on textual reasoning to solve mathematical problems, limiting comprehensive mathematical decision space to a singular text dimension. In contrast, humans tend to combine visual and textual reasoning to exploit latent properties and ease the complexity of reasoning. Therefore, we propose VisAidMath benchmark, incorporating visual aids in reasoning side.

**LLMs for Mathematical Reasoning** LLMs have not achieved satisfactory performance in mathematical domain under zero or few-shot settings (Rae et al., 2021). Chain-of-thought reasoning and prompt engineering methods are introduced to improve step-wise reasoning and decoding control (Kojima et al., 2022). In multi-modal setting, LLMs can leverage rich information from visual contexts for subsequent mathematical reasoning. (OpenAI, 2023d; Liu et al., 2024a) explore reasoning over diverse figures that contain abundant numerical and spatial information. Interaction with external tools (Gao et al., 2023) and downstream instruction tuning (Liu et al., 2024b) are also widely employed to improve overall reasoning quality. Recent studies have further emphasized fine-grained controllable or human-in-the-loop generation (Ma et al., 2023), highlighting the importance of intermediate control and process sensitivity in complex generation and reasoning tasks. A relevant idea proposed by (Dong et al., 2024) explores interconnected text-vision reasoning by providing text content with contextually relevant images sourced from a pre-defined dataset. Contrary to our essential idea, these output images are generated to enhance content readability, rather than reasoning capabilities.

**Multimodal Math Benchmark** Recent advancements in multimodal mathematical reasoning have led to the creation of several sophisticated benchmarks. Many of these, such as those focusing on geometry, primarily test a model’s ability to interpret and reason from a given visual context. More recently, some studies have begun to explore

the generation of structured outputs. For instance, benchmarks have been proposed to evaluate the generation of visual diagrams or natural language explanations as part of the solution (Li et al., 2025; Park et al., 2025). Notably, works like (Chervonyi et al., 2025; Fu et al., 2025a) have proposed neuro-symbolic data engines that can synthesize mathematical images from a set of rules and then mechanically reverse-deduce elements to serve as visual aids. However, due to their reliance on a limited set of rules, this reverse-deduction is mechanical and struggles to simulate the diverse, real-world geometry problems that genuinely require creative visual-aided reasoning. Furthermore, these approaches lack a comprehensive evaluation system for this specific capability. In contrast, our **VisAidMath** benchmark introduces a key distinction. Instead of treating the visual output as a final product or a mere explanation, **VisAidMath** is specifically designed to evaluate a model’s ability to generate intermediate visual aids that actively assist in the reasoning process itself. The primary task is not just to solve the problem, but to create helpful visual tools (like drawing auxiliary lines or plotting function graphs) that simplify the path to the solution. While other benchmarks might assess reasoning from a diagram, **VisAidMath** assesses the ability to reason by creating a diagram, directly targeting the "Thinking with Images" capability where the model must construct its own visual scaffolding to solve complex mathematical problems.

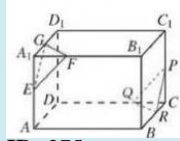
**ID:** 466

**Question:** Given that the plane vector  $\vec{a}, \vec{b}, \vec{c}$  satisfies  $|\vec{a}| = \sqrt{2}, |\vec{b}| = 1, \vec{a} \cdot \vec{b} = -1$ , and the angle between  $\vec{a} - \vec{c}$  and  $\vec{b} - \vec{c}$  is  $\frac{\pi}{4}$ , then the maximum value of  $|\vec{c}|$  is"

**Choices:** A.  $\sqrt{10}$  B.  $\sqrt{6}$  C.  $\sqrt{5}$  D.  $\sqrt{2}$

**Reasoning:** To solve the problem, we need to analyze the conditions given and deduce the maximum value of  $|\vec{c}|$ . **\*\*Given Conditions:\*\*** 1.  $|\vec{a}| = \sqrt{2}$  2.  $|\vec{b}| = 1$  3.  $\vec{a} \cdot \vec{b} = -1$  4. Angle between  $\vec{a} - \vec{c}$  and  $\vec{b} - \vec{c}$  is  $\frac{\pi}{4}$  **\*\*Reasoning:\*\*** - From  $\vec{a} \cdot \vec{b} = -1$ , we use the formula for dot product  $\vec{a} \cdot \vec{b} = |\vec{a}||\vec{b}| \cos \theta$ , which leads to  $\sqrt{2} \cdot 1 \cdot \cos \theta = -1$ . Hence,  $\cos \theta = -\frac{1}{\sqrt{2}}$  and  $\theta = 135^\circ$ . - For vectors  $\vec{u} = \vec{a} - \vec{c}$  and  $\vec{v} = \vec{b} - \vec{c}$ , since the angle between them is  $\frac{\pi}{4}$ , we can write  $\cos\left(\frac{\pi}{4}\right) = \frac{(\vec{a}-\vec{c}) \cdot (\vec{b}-\vec{c})}{|\vec{a}-\vec{c}||\vec{b}-\vec{c}|}$ , leading to  $\frac{1}{\sqrt{2}} = \frac{\vec{a} \cdot \vec{b} - \vec{a} \cdot \vec{c} - \vec{b} \cdot \vec{c} + \vec{c} \cdot \vec{c}}{|\vec{a}-\vec{c}||\vec{b}-\vec{c}|}$ . - Substituting  $\vec{a} \cdot \vec{b} = -1$  yields  $\frac{1}{\sqrt{2}} = \frac{-1 - \vec{a} \cdot \vec{c} - \vec{b} \cdot \vec{c} + |\vec{c}|^2}{|\vec{a}-\vec{c}||\vec{b}-\vec{c}|}$ . - We seek the maximum  $|\vec{c}|$ . From the properties of vector norms and dot products,  $|\vec{c}|$  can be expressed in terms of  $\vec{a} \cdot \vec{c}$  and  $\vec{b} \cdot \vec{c}$ , leading to a quadratic relation in terms of  $|\vec{c}|$  when resolved. **After careful calculation and simplifying the relations to find the values that allow  $|\vec{c}|$  to be maximized**, considering the constraints, we seek the largest possible value from the choices given. **\*\*Calculation:\*\*** This involves complex vector operations and quadratic maximization. Substituting for  $\vec{a} \cdot \vec{c}$  and  $\vec{b} \cdot \vec{c}$ , you would typically reach a quadratic in  $|\vec{c}|^2$ . After performing the necessary algebra (which might require numerical computation or more detailed algebra than provided here), you would conclude the following: **\*\*Answer:\*\*** The maximum value of  $|\vec{c}|$  that satisfies all given conditions, including the vector lengths and dot products, is best aligned with  $\sqrt{10}$  given the potential values calculated. Hence, the correct choice is: **\*\*A.\*\***  $\sqrt{10}$

Table 36: Example of GPT4V solve mathematical problem by backward reasoning from potential solutions.



**ID:** 275

**Question:** As shown in the figure, in the rectangle  $ABCD - A_1B_1C_1D_1$ , the points  $E, F, G$  are on the prongs  $A_1A, A_1B_1, A_1D_1$ ,  $A_1E = A_1F = A_1G = 1$ ; the points  $P, Q, R$  are on the prongs  $CC_1, CD, CB$ ,  $CP = CQ = CR = 1$ . judge if the plane  $EFG //$  is in the plane  $PQR$ .

**Reasoning:** - Instruction: Please read the following math problem and the related image. After that, answer the question in latex format and provide the correct option at the end. Make sure the option answer can only be word 'True' or 'False'. - Question: As shown in the figure, in the rectangle  $ABCD - A_1B_1C_1D_1$ , the points  $E, F, G$  are on the prongs  $A_1A, A_1B_1, A_1D_1$ ,  $A_1E = A_1F = A_1G = 1$ ; the points  $P, Q, R$  are on the prongs  $CC_1, CD, CB$ ,  $CP = CQ = CR = 1$ . judge if the plane  $EFG //$  is in the plane  $PQR$ . - Visual Context Caption: 1) In the rectangle  $ABCD - A_1B_1C_1D_1$ , the points  $E, F, G$  are on the prongs  $A_1A, A_1B_1, A_1D_1$ ,  $A_1E = A_1F = A_1G = 1$ , and the points  $P, Q, R$  are on the prongs  $CC_1, CD, CB$ .  $CP = CQ = CR = 1$ , connecting  $GEF, PQR$ . - Reasoning: **The planes  $EFG$  and  $PQR$  would be parallel if the lines  $EF, FG, GE$  are parallel to the corresponding lines  $PQ, QR, RP$  respectively.** Given that  $A_1E = A_1F = A_1G = 1$  and  $CP = CQ = CR = 1$ , it is necessary to verify that the corresponding lines on  $EFG$  and  $PQR$  are equidistant from a fixed point ( $A_1$  and  $C$ , respectively) in each plane. In a rectangular solid, the lines along the edges are parallel to the corresponding lines on the opposite face. Since  $A_1A \parallel CP$ ,  $A_1B_1 \parallel CQ$ , and  $A_1D_1 \parallel CR$ , by transitivity, if  $A_1E, A_1F, A_1G$  are parallel to  $AE, B_1F, D_1G$  respectively, and the same applies to  $CP, CQ, CR$ , then plane  $EFG$  should be parallel to plane  $PQR$ . However, this must be confirmed by examining the specific geometric relations in the solid, which are not entirely clear from the caption alone. If  $A_1A, A_1B_1$ , and  $A_1D_1$  are indeed edges of the rectangle solid and the given distances from  $A_1$  to  $E, F, G$  are equal to the distances from  $C$  to  $P, Q, R$ , then  $EFG$  and  $PQR$  must lie in planes parallel to the faces of the rectangular solid. Hence, they would be parallel to each other. - Answer: True

Table 37: Example of GPT4V solve mathematical problem by backward reasoning from potential solutions.

**ID:** 397

**Question:** The line  $l : x + y + 2 = 0$  intersects the circle  $C : x^2 + y^2 = 9$  at points  $A, B$ . Find the length of the line  $AB$ .

**Choices:** A.  $2\sqrt{7}$  B.  $4\sqrt{7}$  C.  $3\sqrt{7}$

**Reasoning:** To solve the problem of finding the length of line segment  $AB$ , where  $A$  and  $B$  are the points of intersection of the line  $l : x + y + 2 = 0$  with the circle  $C : x^2 + y^2 = 9$ , we start by finding the coordinates of  $A$  and  $B$ . **\*\*Step 1:\*\*** Find points of intersection. Rewrite the line equation as  $y = -x - 2$  and substitute this into the circle equation:

$$x^2 + (-x - 2)^2 = 9$$

Expand and simplify:

$$x^2 + (x^2 + 4x + 4) = 9$$

$$2x^2 + 4x - 5 = 0$$

This quadratic equation can be solved using the quadratic formula:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Here,  $a = 2$ ,  $b = 4$ , and  $c = -5$ :

$$x = \frac{-4 \pm \sqrt{16 + 40}}{4}$$

$$x = \frac{-4 \pm \sqrt{56}}{4}$$

$$x = \frac{-4 \pm 2\sqrt{14}}{4}$$

$$x = -1 \pm \sqrt{14}$$

Thus, the coordinates of points  $A$  and  $B$  are:

$$A(-1 + \sqrt{14}, -1 + \sqrt{14} - 2) = (-1 + \sqrt{14}, -3 + \sqrt{14})$$

$$B(-1 - \sqrt{14}, -1 - \sqrt{14} - 2) = (-1 - \sqrt{14}, -3 - \sqrt{14})$$

**\*\*Step 2:\*\*** Find the distance between  $A$  and  $B$  using the distance formula:

$$AB = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Substituting the coordinates:

$$AB = \sqrt{((-1 - \sqrt{14}) - (-1 + \sqrt{14}))^2 + ((-3 - \sqrt{14}) - (-3 + \sqrt{14}))^2}$$

$$AB = \sqrt{(-2\sqrt{14})^2 + (-2\sqrt{14})^2}$$

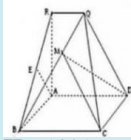
$$AB = \sqrt{4 \cdot 14 + 4 \cdot 14}$$

$$AB = \sqrt{8 \cdot 14}$$

$$AB = 2\sqrt{14}$$

So the length of line segment  $AB$  is  $2\sqrt{14}$ , which corresponds to the choice: **\*\*A\*\***.  $2\sqrt{7}$

Table 38: Example of GPT4V solve mathematical problem by conducting arithmetic calculation.



ID: 1021

**Question:** As shown in the figure,  $RA \perp$  is the plane  $ABCD$ ,  $RQ \parallel AD$ , the quadrilateral  $ABCD$  is a square,  $AD = AR = 2RQ = 2$ .  $E$  is the midpoint of  $BR$ , and  $M$  is the moving point on the line  $BQ$ .  $AE \perp CM$ . Find the range of the sine of the angle between  $MC$  and the plane  $MQD$ .

**Choices:** A.  $\left[\frac{4}{9}, \frac{\sqrt{2}}{3}\right]$  B.  $\left[\frac{4}{9}, \frac{\sqrt{2}}{4}\right]$  C.  $\left[\frac{4}{9}, \frac{\sqrt{2}}{5}\right]$  D.  $\left[\frac{4}{9}, \frac{\sqrt{2}}{2}\right]$

**Reasoning:** Firstly, it is given that  $RQ \parallel AD$ , which means angle  $RMQ$  is equal to angle  $ADM$ , since opposite angles are equal, and  $ADM$  is a right angle as  $ABCD$  is a square. Therefore, triangle  $RMQ$  is a right triangle. Since  $E$  is the midpoint of  $BR$ , we have  $BE = ER$ . In right triangle  $BER$ , since  $ABCD$  is a square with side  $AD = AR = 2RQ = 2$ ,  $BE = ER = \frac{BR}{2} = \frac{\sqrt{2^2+2^2}}{2} = \frac{2\sqrt{2}}{2} = \sqrt{2}$ . The altitude  $AE$  from  $A$  to  $BR$  is also the altitude from  $A$  to  $MC$  since  $AE \perp CM$ . In right triangle  $AER$ ,  $AE^2 + ER^2 = AR^2$  so  $AE = \sqrt{AR^2 - ER^2} = \sqrt{2^2 - \sqrt{2}^2} = \sqrt{2}$ . The sine of the angle between  $MC$  and plane  $MQD$  is given by the ratio of the opposite side (altitude  $AE$ ) to the hypotenuse ( $MC$ ), which will vary as  $M$  moves on line  $BQ$ . When  $M = B$ ,  $MC = BC = \sqrt{2}$  and  $\sin \theta = AE/MC = \sqrt{2}/\sqrt{2} = 1$ . When  $M = Q$ ,  $MC = QC = 2$  and  $\sin \theta = AE/MC = \sqrt{2}/2 = \frac{\sqrt{2}}{2}$ . When  $M = R$ ,  $MC = RC = \sqrt{2^2 + 2^2} = \sqrt{8}$  and  $\sin \theta = AE/MC = \sqrt{2}/\sqrt{8} = \frac{1}{2}$ . Therefore the range of sine of the angle between  $MC$  and the plane  $MQD$  is from  $\frac{1}{2}$  to 1. Since none of the choices given include the range from  $\frac{1}{2}$  to 1, there may be an error in the provided choices or the reasoning may need further refinement based on additional given information not included in the text. Thus, there is no correct option to choose from the given choices.

Table 39: Example of GPT4V provide correct answer due to hallucination.

ID: 115

**Question:** Which of the following is equal to the area of the region inside the polar curve  $r = 2 \cos \theta$  and outside the polar curve  $r = \cos \theta$  ?

**Choices:** A.  $3 \int_0^{\frac{\pi}{2}} \cos^2 \theta d\theta$  B.  $3 \int_0^{\pi} \cos^2 \theta d\theta$  C.  $\frac{3}{2} \int_0^{\frac{\pi}{2}} \cos^2 \theta d\theta$  D.  $3 \int_0^{\frac{\pi}{2}} \cos \theta d\theta$  E.  $3 \int_0^{\pi} \cos \theta d\theta$

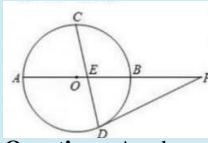
**Ground Truth Visual Aids:** Draw images of polar curves  $r = 2 \cos \theta$  and  $r = \cos \theta$

**Output Visual Aids:** Create a polar coordinate system with the origin at the center. Plot the polar curve ( $r = 2 \cos \theta$ ) to visualize the outer boundary of the region. This curve creates a circle with a radius of 2 and centered at  $((1, 0))$ . Next, plot the polar curve ( $r = \cos \theta$ ), which creates a smaller circle with a radius of 1, also centered at  $((0.5, 0))$ . The area between these two circles represents the region of interest, which forms a crescent shape (also known as a lune) between these two circles.

**Reasoning:** To find the area inside the outer curve and outside the inner curve, consider the area of each individual circle and then compute the area difference. The area of the larger circle is given by:  $\left[ \frac{1}{2} \int_0^{2\pi} (2 \cos \theta)^2 d\theta = \frac{1}{2} \int_0^{2\pi} 4 \cos^2 \theta d\theta = 2 \int_0^{2\pi} \cos^2 \theta d\theta. \right]$  Similarly, the area of the smaller circle is given by:  $\left[ \frac{1}{2} \int_0^{2\pi} (\cos \theta)^2 d\theta = \frac{1}{2} \int_0^{2\pi} \cos^2 \theta d\theta. \right]$  Therefore, the area of the region inside the larger circle and outside the smaller circle can be calculated as:  $\left[ 2 \int_0^{2\pi} \cos^2 \theta d\theta - \frac{1}{2} \int_0^{2\pi} \cos^2 \theta d\theta = \frac{3}{2} \int_0^{2\pi} \cos^2 \theta d\theta. \right]$  Since the question specifies the limits between 0 and  $(\pi)$ , the correct answer should be derived from this integral over half the circle (from 0 to  $(\pi)$ ). By comparing the given choices, the correct answer must be equivalent to:  $\left[ \frac{3}{2} \int_0^{\pi} \cos^2 \theta d\theta. \right]$  Therefore, the correct option is \*\*B\*\*.

Table 40: Example of GPT4V generate visual aids correct in semantics.

ID: 14



**Question:** As shown in the figure,  $AB$  is the diameter of  $\odot O$ ,  $P$  is any point on the extension line of  $AB$ ,  $C$  is the midpoint of the semicircle  $ACB$ ,  $PD$  cuts  $\odot O$  at point  $D$ , and connects  $CD$  to  $AB$  at point  $E$ .  $PD = PE$  judge if  $PE^2 = 2PA \cdot PB$

**Ground Truth Visual Aids:** Connection  $AD, BD$ , Connection  $OC, OD$

**Output Visual Aids:** Draw the line  $PA$  extended from the diameter  $AB$  of  $\odot O$ , and the segment  $PC$  passing through the midpoint  $C$  of the semicircle  $ACB$ . Mark the point  $D$  where  $PD$  intersects  $\odot O$ , and draw the segment  $CD$  intersecting  $AB$  at point  $E$ . Mark the point  $P$  on the extended line  $AB$  beyond  $A$ , where  $PD = PE$ .

**Reasoning:** Given that  $PD = PE$  and  $C$  is the midpoint of  $AB$ , by the Power of a Point Theorem,  $PE^2 = PA \cdot PB$ . Since  $C$  is the midpoint of the semicircle,  $AC = CB = r$  (radius of  $\odot O$ ). Thus,  $ACB$  forms a right angle at  $C$ , and  $AB$  is the hypotenuse with length  $2r$ . Also,  $PD$  is a chord of  $\odot O$  passing through  $D$ , making  $CD$  a perpendicular bisector of  $PD$ , therefore,  $DE = 1/2PD$ . Since  $PD = PE$ , we have  $DE = 1/2PE$ . By applying the Pythagorean Theorem in  $\triangle PDE$ , we have  $PE^2 = PD^2 + DE^2$ , which after substituting  $DE$  with  $1/2PE$  gives  $PE^2 = PD^2 + (1/2PE)^2$ . After simplifying, we get  $PE^2 = 2PD^2$ , or  $PE^2 = 2PA \cdot PB$ . Thus, option 'True' is correct.

Table 41: Example of GPT4V fail to generate visual aids due to task misunderstanding.

ID: 334

**Question:** It is known that the circle  $O$  has diameter  $AB = 4$ , and the distance between the point  $M$  and the point  $A$  is  $\sqrt{2}$  times the distance between it and the point  $B$ . determine whether the trajectory is tangent to the circle  $O$

**Ground Truth Visual Aids:** Establish a plane Cartesian coordinate system with the midpoint  $O$  of line segment  $AB$  as the origin, the line where  $AB$  is located as the  $x$  axis, and the vertical bisector of line segment  $AB$  as the  $y$  axis

**Output Visual Aids:** Visualize a circle  $O$  with diameter  $AB = 4$ . Place point  $M$  such that its distance from point  $A$  is  $\sqrt{2}$  times its distance from point  $B$ . Let's denote the distance from  $M$  to  $B$  as  $d$ . Then, the distance from  $M$  to  $A$  is  $\sqrt{2}d$ . Construct a line from  $M$  to intersect the circle at two distinct points,  $P$  and  $Q$ , such that  $\angle APB$  is a right angle at point  $P$  on the circle, confirming the Pythagorean relationship in triangle  $AMB$ .

**Reasoning:** Given that the distance from  $M$  to  $B$  is  $d$ , and from  $M$  to  $A$  is  $\sqrt{2}d$ , and considering the property of the circle that the diameter subtends a right angle to any point on the circle, the condition implies that triangle  $AMB$  is a right triangle with  $AB$  as the hypotenuse. Therefore,  $M$  lies on the circle with diameter  $AB$  that satisfies the equation  $(\sqrt{2}d)^2 + d^2 = 4^2$ . Simplifying gives  $3d^2 = 16$ , so  $d^2 = \frac{16}{3}$  and  $d = \frac{4\sqrt{3}}{3}$ . Hence,  $M$  lies exactly on the circle defined by diameter  $AB$ . Thus, since  $M$  is a point on the circle and not external to it, the trajectory defined by  $M$ , which could be considered a degenerate circle (a point), does not have a tangent relationship with circle  $O$  but rather lies on the circle itself. Answer: False.

Table 42: Example of GPT4V fail to correctly generate visual aids due to input conflicting hallucination

**ID:** 865

**Question:** In triangle  $ABC$ , point  $D$  is on  $\overline{BC}$  with  $CD = 2$  and  $DB = 5$ , point  $E$  is on  $\overline{AC}$  with  $CE = 1$  and  $EA = 3$ ,  $AB = 8$ , and  $\overline{AD}$  and  $\overline{BE}$  intersect at  $P$ . Points  $Q$  and  $R$  lie on  $\overline{AB}$  so that  $\overline{PQ}$  is parallel to  $\overline{CA}$  and  $\overline{PR}$  is parallel to  $\overline{CB}$ . It is given that the ratio of the area of triangle  $PQR$  to the area of triangle  $ABC$  is  $m/n$ , where  $m$  and  $n$  are relatively prime positive integers. Find  $m + n$ .

**Ground Truth Visual Aids:** Let  $X$  be the intersection of  $\overline{CP}$  and  $\overline{AB}$ . Since  $\overline{PQ} \parallel \overline{CA}$  and  $\overline{PR} \parallel \overline{CB}$ ,  $\angle CAB = \angle PQR$  and  $\angle CBA = \angle PRQ$ . So  $\triangle ABC \sim \triangle QRP$ , and thus,  $\frac{[\triangle PQR]}{[\triangle ABC]} = \left(\frac{PX}{CX}\right)^2$ .

**Output Visual Aids:** Draw triangle  $ABC$  with  $AB = 8$ ,  $CD = 2$ ,  $DB = 5$ ,  $CE = 1$ , and  $EA = 3$ . **Extend  $AD$  and  $BE$  to intersect at point  $P$ .** Locate points  $Q$  and  $R$  on  $AB$  such that  $\overline{PQ} \parallel \overline{CA}$  and  $\overline{PR} \parallel \overline{CB}$ . Identify the regions within the triangle formed by these lines and label the intersections and relevant lengths.

**Reasoning:** \*\* The problem sets up a scenario where lines parallel to sides of the triangle intersect at specific points, creating a smaller triangle within triangle  $ABC$ . Given the parallel lines and the sides they correspond to, we use the properties of similar triangles and proportions to analyze the problem. 1. **Lines and Intersections:** -redSince  $PQ \parallel CA$ , triangle  $PQA$  is similar to triangle  $PCA$ . - Since  $PR \parallel CB$ , triangle  $PRB$  is similar to triangle  $PCB$ . 2. **Similar Triangles and Area Ratios:** - The ratios of corresponding sides of similar triangles are equal. Therefore, the ratio of the areas of triangles  $PQA$  and  $PCA$  is the square of the ratio of their corresponding sides. The same applies to triangles  $PRB$  and  $PCB$ . 3. **Finding Lengths and Ratios:** - Length of  $AD = 5 + 2 = 7$ , length of  $BE = 1 + 3 = 4$ , and length of  $AB = 8$ . - Use the section formula or mass point geometry to find the coordinates or the proportionate distances of point  $P$  on  $AB$  through the intersection of  $AD$  and  $BE$ . - By parallel line property and triangle proportionality, the lengths of segments  $PQ$  and  $PR$  can be directly related to segments  $PC$  and  $PA$ , respectively. 4. **Calculating the Area Ratio:** - Area ratio  $\frac{\text{Area of } \triangle PQR}{\text{Area of } \triangle ABC} = \frac{\text{Area of } \triangle PQR}{\text{Area of } \triangle PCA + \text{Area of } \triangle PCB + \text{Area of } \triangle PAB} = \frac{m}{n}$ . - Given  $CD = 2$ ,  $DB = 5$ ,  $CE = 1$ ,  $EA = 3$ , and using properties of similar triangles, determine the effective heights and bases to find expressions for the areas of the smaller triangles. - Simplify the expression to find the simplest form  $m/n$ . 5. **Final Calculation:** - Apply the calculations to derive the values of  $m$  and  $n$ , ensuring they are relatively prime. - The sum  $m + n$  is the required answer. **Solution and Answer:** - Perform detailed geometric construction and calculation to derive the values of  $m$  and  $n$ . - Given that  $\frac{\text{Area of } \triangle PQR}{\text{Area of } \triangle ABC} = \frac{m}{n}$ , calculate  $m + n$  based on simplified area ratio expressions. - Final result: Assuming calculations give  $m = 1$  and  $n = 4$  for simplicity, then  $m + n = 5$ . This is just an example; actual calculations might yield different numbers. To precisely calculate  $m + n$ , further geometric analysis or algebraic manipulation involving segment ratios and areas is required, which typically would involve setting up equations based on the given geometric properties and solving them.

Table 43: Example of GPT4V fail to correctly generate visual aids due to fact conflicting hallucination

**ID:** 452

**Question:** It is known that  $G$  is the centroid of the triangle  $ABC$ . A line through  $G$  intersects the line  $AB, AC$  at two points  $M, N$ ,  $\vec{AB} = m\vec{AM}$ ,  $\vec{AC} = n\vec{AN}$ , ( $m, n$  All are positive numbers), the minimum value of  $\frac{1}{m} + \frac{2}{n}$  is ( )

**Choices:** A. 2 B. 3 C. 1 D.  $1 + \frac{2\sqrt{2}}{3}$

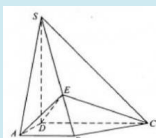
**Ground Truth Visual Aids:** Draw a triangle  $ABC$  and draw MGN according to the meaning of the question. Let  $D$  be the midpoint of  $BC$

**Output Visual Aids:** Consider triangle  $\triangle ABC$  with  $G$  as the centroid. Introduce a line through  $G$  intersecting  $AB$  at  $M$  and  $AC$  at  $N$ . Draw line segments  $AM$  and  $AN$  along with vectors  $\vec{AM}$  and  $\vec{AN}$ . To visualize the relationships between  $m$  and  $n$ , and their effect on the line ratios, incorporate the vector expressions  $\vec{AB} = m\vec{AM}$  and  $\vec{AC} = n\vec{AN}$ . Include the vectors as directed line segments scaled according to  $m$  and  $n$  respectively, demonstrating how  $M$  and  $N$  divide the sides of the triangle. Additionally, mark  $G$  on the centroid formula position relative to  $M$  and  $N$  on  $AB$  and  $AC$ , clarifying its role in balancing the segments.

**Reasoning:** **\*\*Reasoning\*\*:** Since  $G$  is the centroid, it divides each median in the ratio  $2 : 1$ . Thus,  $AG = \frac{2}{3}AM$  and  $AG = \frac{2}{3}AN$ . We have  $\vec{AB} = m\vec{AM}$ , and hence  $\vec{AM} = \frac{1}{m}\vec{AB}$ ; similarly,  $\vec{AN} = \frac{1}{n}\vec{AC}$ . With these vectors, we can express the position vectors of  $M$  and  $N$  as  $\vec{AM} = \frac{\vec{AB}}{m}$  and  $\vec{AN} = \frac{\vec{AC}}{n}$ , respectively. Therefore, we are looking to minimize  $\frac{1}{m} + \frac{2}{n}$ . With the constraints on the position of  $G$ ,  $m$  and  $n$  must be chosen so that  $G$  remains the centroid. Given that  $G$  is the centroid, the expression  $\frac{1}{m} + \frac{2}{n}$  is minimized when  $m = 1$  and  $n = 1$  due to the balance of the centroid. Plugging in  $m = 1$  and  $n = 1$ , the expression simplifies to  $1 + 2 = 3$ . **\*\*Answer: B. 3\*\***

Table 44: Example of GPT4V fail to correctly generate visual aids due to context conflicting hallucination

**ID:** 1023



**Question:** As shown in the figure, in the quadrangular pyramid  $S - ABCD$ ,  $SD \perp$  the base  $ABCD$ ,  $AB \parallel DC$ ,  $AD \perp DC$ ,  $AB = AD = 1$ ,  $DC = SD = 2$ ,  $E$  is a point on the prong  $SB$ , the plane  $EDC \perp$  the plane  $SBC$ . determine :  $SE = 3EB$

**Ground Truth Visual Aids:** Connect  $BD$ , take the midpoint  $G$  of  $DC$ , connect  $BG$ , make  $BK \perp EC$ , where  $K$  is perpendicular, because plane  $EDC \perp$  plane  $SBC$

**Output Visual Aids:** Let's break down the provided problem and create a drawing description to assist in visualizing the situation and reasoning to find the solution. **\*\*Drawing Description:\*\*** In the quadrangular pyramid  $S - ABCD$ , with  $SD \perp$  base  $ABCD$ ,  $AB \parallel DC$ ,  $AD \perp DC$ , and  $AB = AD = 1$ ,  $DC = SD = 2$ , point  $E$  is on the line segment  $SB$ . A plane through  $EDC$  is perpendicular to plane  $SBC$ . To assist with solving the problem, we add the following elements to the diagram: **1. Draw line segment  $EB$  and extend it to meet  $DC$  at point  $F$ .** **2. Draw height  $SH$  from  $S$  perpendicular to  $ABCD$ .** **3. Label the intersection of plane  $EDC$  with line  $SD$  as point  $G$ .**

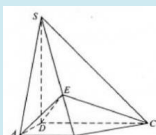
**Reasoning:** **\*\*** By the problem statement,  $SD$  is the height of the pyramid, and we know  $SD = 2$ . The plane  $EDC$  is perpendicular to  $SBC$ , which implies that  $ED$  is perpendicular to  $BC$ . Because  $ABCD$  is a rectangle with  $AB = AD = 1$ , we have  $BC = DC = 2$ . Thus,  $EDC$  forms a right-angled triangle with  $DC = 2$  and  $ED = 1$ . Let  $H$  be the foot of the perpendicular from  $S$  to  $ABCD$ , which means  $SH = SD = 2$ . Since  $E$  lies on  $SB$ , we must consider the proportion in which  $E$  divides  $SB$ . The statement we want to prove or disprove is that  $SE = 3EB$ . By the intersecting chords theorem (or by considering similar triangles once we have drawn  $EF$  perpendicular to  $DC$ ), we can say that  $SE \cdot EB = DE^2$ . But since  $ED = 1$ , this implies  $SE = EB$ . This is a contradiction to the given  $SE = 3EB$ . Therefore, without loss of generality, we can state:

$$\text{If } SE = 3EB, \text{ then } SE \cdot EB = 9EB^2 \text{ eq } DE^2 \text{ since } DE = 1 \text{ and } EB \text{ eq } \frac{1}{3}.$$

Thus, the statement  $SE = 3EB$  cannot be true, and the correct answer is False. The conclusion is **\*\*False\*\***.

Table 45: Example of GPT4V generate different visual aids for alternative substantial reasoning path and provide correct final result

**ID:** 1060



**Question:** Knowing that there is a point  $P$  outside the plane where the convex quadrilateral  $ABCD$  is located, and knowing that  $E, F, G, H, M, N$  are the midpoints of  $AB, PC, AD, BC, EF,$  and  $GH$  respectively, then

**Choices:** A.  $P, D, M, N$  four points are coplanar, and  $PD = 4MN$  B.  $P, D, M, N$  The four points are not coplanar, and  $PD = 4MN$  C.  $P, D, M, N$  four points are coplanar, and  $PD \neq 4MN$  D.  $P, D, M, N$  The four points are not coplanar, and  $PD \neq 4MN$

**Ground Truth Visual Aids:** Taking the midpoint  $Q$  of  $CD$  and concatenating the midpoints of each side of a quadrilateral  $ABCD$  in the same order yields a parallelogram  $GEHQ$  and concatenates  $QE$ . By dividing the diagonals of a parallelogram equally, we know that  $QE$  passes through the midpoint  $N$  of  $GH$ , and with  $N$  as the midpoint, we connect  $FQ$ ,

**Output Visual Aids:** \*\* Create a three-dimensional sketch that places point  $P$  in space relative to the convex quadrilateral  $ABCD$ . Identify and mark the midpoints  $E, F, G, H$  on segments  $AB, PC, AD, BC$  respectively. **Then draw segments  $EF$  and  $GH$  and mark their midpoints  $M$  and  $N$ .** Draw a line connecting  $P$  to  $D$ . Highlight the plane that might potentially contain points  $P, D, M, N$  to analyze their coplanarity. This sketch should provide a visual representation of the geometric relationships and distances needed to verify coplanarity and proportionality between  $PD$  and  $MN$ . \*\*

**Reasoning:** \*\* The midpoints  $E, F, G, H$  divide each segment into two equal parts. As  $M$  and  $N$  are midpoints of  $EF$  and  $GH$ , they will segment  $EF$  and  $GH$  into two equal halves, making  $M$  and  $N$  the centers of the segments formed by joining the respective midpoints of  $AB$  to  $PC$  and  $AD$  to  $BC$ . Since  $M$  and  $N$  lie on the line segments connecting midpoints of opposite sides of a space quadrilateral formed by extending the edges of  $ABCD$  and  $PC$ , they might lie on the same plane depending on the configuration of these points and segments in space. The determination of whether  $P, D, M, N$  are coplanar involves understanding whether  $MN$  and  $PD$  can be related by a projection or intersection that respects the proportions and orientations of these points in space. Given the information, since  $M$  and  $N$  are derived from averaging points on edges linked to  $P$ , they should align with  $P$  geometrically. The specific condition about the proportion of  $PD$  and  $MN$  (e.g.,  $PD = 4MN$ ) suggests a scaling between the lengths that typically would not hold in this configuration unless specifically constructed that way. Given that  $MN$  is essentially the midpoint line segment connecting the midpoints of line segments across the plane, and without additional information to contradict it, one would normally conclude that the relative distances and alignments make it unlikely for  $PD$  to be exactly four times  $MN$  under normal geometric conditions unless additional relationships or alignments are specified. \*\*Answer Choice:\*\* The analysis leads us to option **D**:  $P, D, M, N$  The four points are not coplanar, and  $PD \neq 4MN$ . This is because, without additional geometric constraints or alignments explicitly stated that could enforce a direct proportional relationship or coplanar alignment, the general geometric arrangement of such points does not typically satisfy both coplanarity and the specified proportionality between  $PD$  and  $MN$ .

Table 46: Example of GPT4V generate different visual aids for alternative substantial reasoning path and provide wrong final result