VisAidMath: Benchmarking Visual-Aided Mathematical Reasoning

Jingkun Ma[1] **Runzhe Zhan**[1] **Yang Li**[1] **Di Sun**[2]
Hou Pong Chan[3] **Lidia S. Chao**[1] **Derek F. Wong**[1]*
[1]NLP[2]CT Lab, Department of Computer and Information Science, University of Macau
[2] University of Macau [3]DAMO Academy, Alibaba Group
nlp2ct.{jingkun, runzhe, liyang}@gmail.com, yc47479@um.edu.mo
houpong.chan@alibaba-inc.com, {lidiasc, derekfw}@um.edu.mo

Homepage: https://nlp2ct.github.io/VisAidMathHomepage/
Evaluation: https://www.codabench.org/competitions/7634/

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 host an evaluation platform at CodaBench for testing publicly.

1 Introduction

Mathematical problem-solving (MPS) remains a significant hurdle for Large Language Models (LLMs) and Large Multi-modal Models (LMMs) [1, 2, 3, 4]. The complexity intensifies when problems are presented in a multi-modal format [5, 6]. While some research has explored multi-modal MPS by incorporating visual contexts [6, 7], 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 [8].

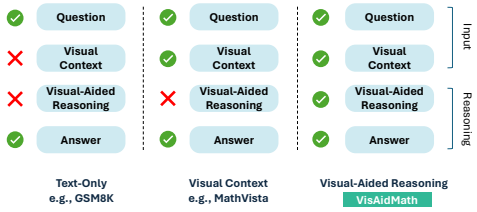
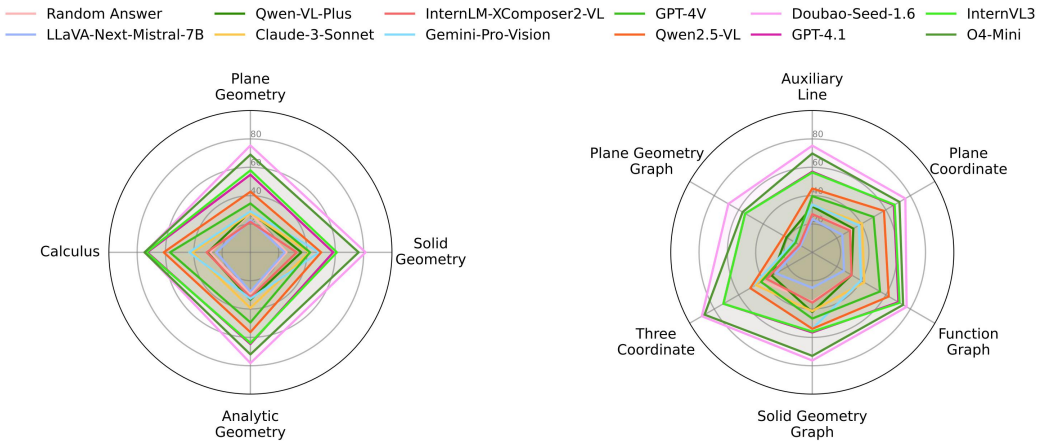


Figure 1: Comparison between VisAidMath and other benchmarks. Our work particularly focuses on utilization of explicit and implicit visual context during reasoning process.

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: VLM4RWD.



(a) Mathematical Branch

(b) Visual Aid

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

This creates a significant gap between evaluating surface-level "Think about Images" and "Think with Images" [8], 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 [8]. Problems within our benchmark can be explored through the selection of analytical tools [9, 10, 11, 12], the programmatic creation of visual aids like auxiliary lines [13, 14, 15, 16], or by engaging in intrinsic spatial imagination to foresee the solution path [17, 18, 19].

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. 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 utilizing 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 J. 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.

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. A comprehensive description of our data sources, collection methodology, metadata, and translation process is detailed in 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.

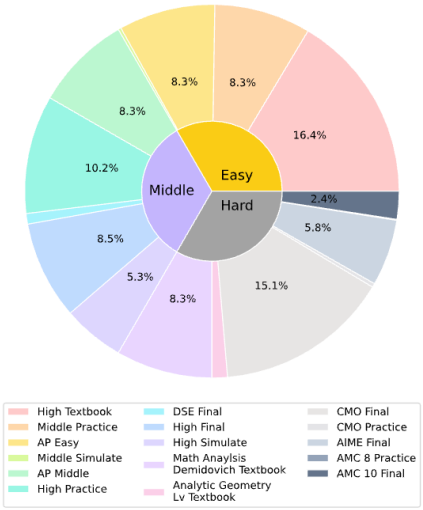


Figure 3: Distribution of data sources and difficulty levels.

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 2.2). 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. 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 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 LaTeX files using OCR.
- `Annotator`: The annotator first validates and refines the LaTeX files by comparing the original PDF files provided by the collector with the transformed LaTeX 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.

2.3 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

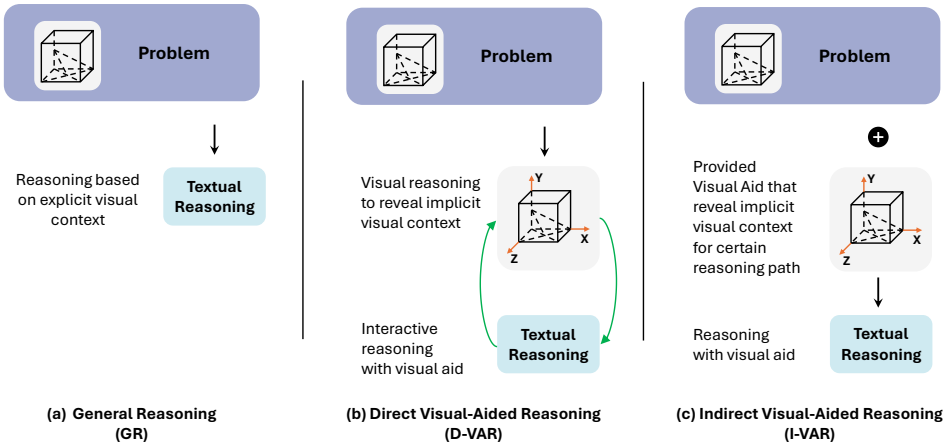


Figure 4: Comparison of different tasks: a) General Reasoning: provide MPS reasoning steps directly. b) Direct Visual-Aided Reasoning: create visual aids that disclose implicit visual context within problem, incorporating with textual reasoning to solve mathematical problem. c) Indirect Visual-Aided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

distribution. This balance enables a broader exploration of diverse mathematical knowledge. Further details on other attributes can be found in Appendix C.

2.4 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.

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	$(C, Q) \rightarrow (V_g, A)$	**Generates** a visual aid to assist reasoning.
Indirect Visual-Aided Reasoning	I-VAR	$(C, Q, V_p) \rightarrow A$	**Utilizes** a provided visual aid for reasoning.

To accommodate language models that only accept textual inputs, the visual context ‘C‘ is replaced by its textual description, denoted as C_{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, 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.

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.

Model	ALL	PLG	SDG	AYG	CAL	AXL	RTC	THC	PLG	SDG	FUG
Heuristics Baselines											
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
Large Language Models (LLMs): Text-Only Input											
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
Large Multimodal Models (LMMs): Text-Only Input											
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
Large Multimodal Models (LMMs): Multimodal Input											
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

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 [20], Mistral-7B-Instruct-v0.2 [21]; 2) Closed-source LLMs: GPT-3.5-turbo [22], GPT-4-turbo [23]; 3) Open-source LMMs: LLaVA-Next-Mistral-7B [24], InternLM-XComposer2-VL [25], VL-Cogito[26], Qwen2.5-VL-72B[27], InternVL3.5-38B[28]; 4) Closed-source LMMs: Qwen-VL-Plus [29], Gemini-Pro-Vision [30], Claude-3-Sonnet [31], GPT-4-Vision [32], GPT-4.1[33], o4-mini[34], and Doubao-Seed-1.6[35]. 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 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

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 **PVA**, and **SPRS**. 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**.

Model	ACCU	PVA	SPRS	LogiR	ToolF	InfoT	GeneR
Task 1: General Reasoning							
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
Task 2: Direct Visual-Aided Reasoning							
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
Task 3: Indirect Visual-Aided Reasoning							
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

answers, but also robustly derived solutions from those that are superficially correct but procedurally flawed.

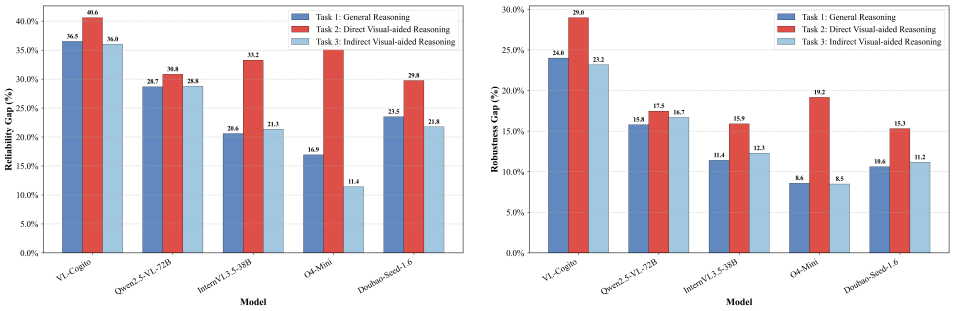
Final Answer Correctness (ACCU) Given that VisAidMath comprises mathematics problems with deterministic answers, we evaluate the correctness of the final answer using Standard Accuracy (ACCU) as a baseline metric. To accurately extract the final answer from model outputs, we follow the approach of [7] 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 section F.5. ACCU is calculated as:

$$\text{ACCU} = \frac{1}{N} \sum_{j=1}^{N} a_j \tag{1}$$

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 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 ($\mathcal{D}$ = LogiR, ToolF, InfoT, GeneR). These dimensions are defined as follows:



(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.

- **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.

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:

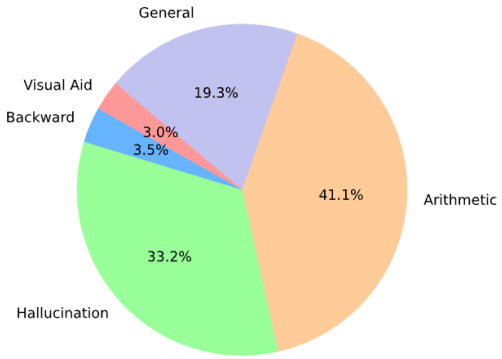
$$\text{PVA} = \frac{1}{N} \sum_{j=1}^{N} a_j \cdot V_j \tag{2}$$

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.

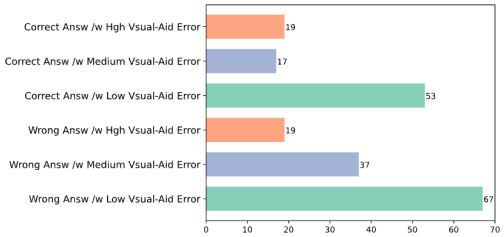
Solution Process Robustness Score (SPRS) The framework culminates in the Solution Process Robustness Score (SPRS), which offers the most granular level of assessment. Distinct from the binary filtering of PVA, SPRS provides a continuous, fine-grained quality score for all correctly answered problems. 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 calculated as:

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

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.



(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.

3.3 Main Results

Initial Performance via Standard Accuracy The complete performance results are presented in Tables 28, 2, and 29, with a summary of leading 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.

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 *evade* the intended visual-aided path. Our manual 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.

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.

References

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

[2] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving

quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.

[3] Qinzhuo Wu, Qi Zhang, Jinlan Fu, and Xuan-Jing Huang. 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, 2020.

[4] Yuhuai Wu, Markus N Rabe, Wenda Li, Jimmy Ba, Roger B Grosse, and Christian Szegedy. Lime: Learning inductive bias for primitives of mathematical reasoning. In *International Conference on Machine Learning*, pages 11251–11262. PMLR, 2021.

[5] Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022.

[6] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

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

[8] Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025.

[9] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*, 2022.

[10] Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind's eye of llms: visualization-of-thought elicits spatial reasoning in large language models. *Advances in Neural Information Processing Systems*, 37:90277–90317, 2024.

[11] Dairu Liu, Ziyue Wang, Minyuan Ruan, Fuwen Luo, Chi Chen, Peng Li, and Yang Liu. Visual abstract thinking empowers multimodal reasoning. *arXiv preprint arXiv:2505.20164*, 2025.

[12] Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, et al. Cogcom: A visual language model with chain-of-manipulations reasoning. *arXiv preprint arXiv:2402.04236*, 2024.

[13] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14953–14962, 2023.

[14] Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. *arXiv preprint arXiv:2501.05452*, 2025.

[15] Yuri Chervonyi, Trieu H Trinh, Miroslav Olšák, Xiaomeng Yang, Hoang Nguyen, Marcelo Menegali, Junehyuk Jung, Vikas Verma, Quoc V Le, and Thang Luong. Gold-medalist performance in solving olympiad geometry with alphageometry2. *arXiv preprint arXiv:2502.03544*, 2025.

[16] Shichao Weng, Zhiqiang Wang, Yuhua Zhou, Rui Lu, Ting Liu, Zhiyang Teng, Xiaozhang Liu, and Hanmeng Liu. Geosketch: A neural-symbolic approach to geometric multimodal reasoning with auxiliary line construction and affine transformation. *arXiv preprint arXiv:2509.22460*, 2025.

[17] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[18] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Rui Huang, Haoquan Zhang, Manyuan Zhang, Jiaming Liu, Shanghang Zhang, Peng Gao, et al. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025.

[19] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. 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, 2025.

[20] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[21] Gueyoung Jung, Matti A Hiltunen, Kaustubh R Joshi, Richard D Schlichting, and Calton Pu. 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, 2010.

[22] OpenAI. https://platform.openai.com/docs/models/gpt-3-5-turbo. In *OpenAI*. OpenAI, 2023.

[23] OpenAI. https://platform.openai.com/docs/models/gpt-4. In *OpenAI*. OpenAI, 2023.

[24] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.

[25] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.

[26] Ruifeng Yuan, Chenghao Xiao, Sicong Leng, Jianyu Wang, Long Li, Weiwen Xu, Hou Pong Chan, Deli Zhao, Tingyang Xu, Zhongyu Wei, et al. Vl-cogito: Progressive curriculum reinforcement learning for advanced multimodal reasoning. *arXiv preprint arXiv:2507.22607*, 2025.

[27] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[28] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[29] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

[30] Google. https://labelbox.com/product/model/foundry-models/google-gemini-pro-vision/. In *Google*. Google, 2023.

[31] AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 2024.

[32] OpenAI. https://openai.com/index/gpt-4v-system-card/. In *OpenAI*. OpenAI, 2023.

[33] OpenAI. https://platform.openai.com/docs/models/gpt-4.1. In *OpenAI*. OpenAI, 2023.

[34] OpenAI. https://platform.openai.com/docs/models/o4-mini. In *OpenAI*. OpenAI, 2023.

[35] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

[36] larryflynt. https://github.com/larryflynt/image-concat. In *Github*, 2023.

[37] 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. Transformers: State-of-the-art

natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[38] Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. How well do large language models perform in arithmetic tasks? *arXiv preprint arXiv:2304.02015*, 2023.

[39] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023.

[40] Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746*, 2022.

[41] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021.

[42] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

[43] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[44] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR, 2023.

[45] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[46] Can Li, Ying Liu, Ting Zhang, Mei Wang, and Hua Huang. Visiomath: Benchmarking figure-based mathematical reasoning in lmms. *arXiv preprint arXiv:2506.06727*, 2025.

[47] Jaewoo Park, Jungyang Park, Dongju Jang, Jiwan Chung, Byungwoo Yoo, Jaewoo Shin, Seonjoon Park, Taehyeong Kim, and Youngjae Yu. Explain with visual keypoints like a real mentor! a benchmark for multimodal solution explanation. *arXiv preprint arXiv:2504.03197*, 2025.

[48] Daocheng Fu, Zijun Chen, Renqiu Xia, Qi Liu, Yuan Feng, Hongbin Zhou, Renrui Zhang, Shiyang Feng, Peng Gao, Junchi Yan, et al. Trustgeogen: Scalable and formal-verified data engine for trustworthy multi-modal geometric problem solving. *arXiv preprint arXiv:2504.15780*, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims has been made in the abstract and introduction accurately.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitation is discueseed in appendix K.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: There is no theoretical result provided in this work.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Detailed information for reproducing the main experimental results is provided in section 2, and appendix C.1 and E.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our data will be hosted on an evaluation platform at CodaBench for testing publicly.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Training and test details are provided in appendix F.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The experiments were each conducted only a single time due to expenses limit. As a result, it was not possible to calculate error bars or determine the statistical significance of the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Information of computer resources used in this work is hard to record.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work is conducted conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Social impact is discussed in appendix K.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: This paper properly cited and credited the paper of code, data, and models used in this work.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Detail of data liscence is provided in appendix B.3

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The detail of compensation and screenshots are provided in appendix E.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Refer to appendix E.1.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Refer to appendix L.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix

Contents

A Future Direction **22**

B Dataset Preparation **22**

B.1 Machine Translation 22

B.2 Data Processing 22

B.3 Formalization 22

C Dataset Analysis **23**

C.1 Metadata 23

C.2 Data Source 23

D Examples for Different Categorizations **23**

D.1 Math Branch 23

D.2 Visual Aid Type 24

D.3 Complexity 24

E Dataset Collection Detail **24**

E.1 Annotation Details 24

E.2 Caption Writing Templates 26

E.3 Dataset Creation Pipeline 26

E.4 Human Annotation Interface 26

F Detail Experiment Settings **33**

F.1 Hyperparameters 33

F.2 Reasoning Prompter 33

F.3 Instruction across Answer Types 35

F.4 Visual Aid Extraction 35

F.5 Answer Extraction Prompter 35

G More Experimental Results **35**

G.1 Results of other tasks 35

G.2 Quantitative Analysis 35

G.2.1 Performance across Math Branches 36

G.2.2 Performance across Complexities 39

G.2.3 Performance across Visual Aids 39

G.3 N-gram Similarities 40

H In-depth Model Analysis **49**

H.1 General Reasoning Tendency 49

H.2 Visual Aid Inference Capability 49

I Supplementary Quantitative Analysis **49**

J Related Work **52**

K limitation and social impact **56**

L LLM Usage Declaration **56**

A Future Direction

Spatial Capability Despite the predominant emphasis on the construction and fitting of extensive datasets, mainstream works are confined to inference tasks within textual dimension. LLM exhibits exceedingly poor performance in providing visual reasoning step, revealing deficiencies in spatial understanding, imagination, and more other aspects. To address mathematical reasoning with visual-aid inference, future investigation could be directed to specifically enhance LLM's adaptation to visual reasoning tasks, devise new methods for better integration of textual and visual reasoning, or design specific architectures for multimodal reasoning steps 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[2] and DeepL[3] 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.

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 [36] 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.

[2] `https://fanyi.baidu.com/`

[3] `https://www.deepl.com/translator`

Table 4: Detail metadata for **VisAidMath**

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

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 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, discuessed 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, 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.

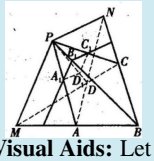
Table 5: Detail of data sources

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 6: Example of mathematical problems of plane geometry.

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

D.2 Visual Aid Type

D.3 Complexity

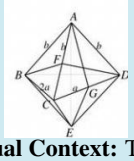
E Dataset Collection Detail

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 LaTeX file to ensure completeness and accuracy. Following the approach used in MathVista [7], 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

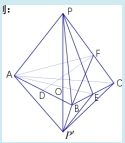
Table 7: Example of mathematical problems of solid geometry.

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.

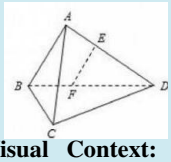
Question: Given that two congruent triangular pheons are glued together to obtain a hexahedron with all the dihedral angles equal, and that the shortest prong of the hexahedron is 2, the distance between the two farthest vertices is

Visual Aids: 

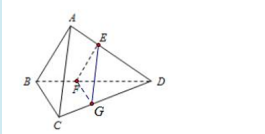
Caption of Visual Aids: Make a plane $BDF \perp$ with edges AC intersecting AC at F

Table 8: Example of mathematical problems of analytic geometry.

ID: 619

Visual Context: 

Caption of Visual Context: In the trigonometry $\mathrm{A}-\mathrm{BCD}$, $\mathrm{AB} \perp \mathrm{AD}, \mathrm{BC} \perp \mathrm{BD}$, the plane $\mathrm{ABD} \perp$ the plane BCD, points $\mathrm{E}, \mathrm{F}(\mathrm{E}$ do not coincide with $\mathrm{A}, \mathrm{D})$ are on the prisms AD, BD and EF BD respectively. $\perp$ AD . EF / $/Plane$ ABC
Question: In the trigonal pheasant $\mathrm{A}-\mathrm{BCD}$, $\mathrm{AB} \perp \mathrm{AD}, \mathrm{BC} \perp \mathrm{BD}$, plane $\mathrm{ABD} \perp$ plane BCD , points $\mathrm{E}, \mathrm{F}(\mathrm{E}$ and A, D do not coincide) on the prong AD, BD respectively, and $\mathrm{EF} \perp \mathrm{AD}.\mathrm{EF}//$ plane ABC determine whether $\mathrm{AD} \perp \mathrm{AC}$.

Visual Aids: 

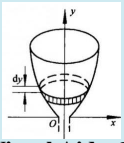
Caption of Visual Aids: Take a point Gon the line segment CD, connect FG, EGso that $\mathrm{FG}//\mathrm{BC}$

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. The task was described as minimal risk.

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

ID: 939

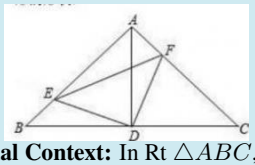
Question: Can the surface of a container formed by rotating the curve $y = Cx^4$about the vertical axis Oycause the liquid surface to fall uniformly as it flows from the bottom of the container?

Visual Aids: 

Caption of Visual Aids: In a geometric body obtained by rotating an upward parabola along the xaxis, dxis a line segment parallel to the xaxis, and the volume enclosed by it is drawn

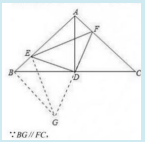
Table 10: Example of mathematical problem with auxiliary line as visual aid.

ID: 10

Visual Context: 

Caption of Visual Context: In Rt $\triangle ABC$, $\triangle BAC = 90^\circ$, 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 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.

Visual Aids: 

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

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 Dataset Creation Pipeline

The pipeline is shown in the Figure 9.

E.4 Human Annotation Interface

Shown in the Figure 10.

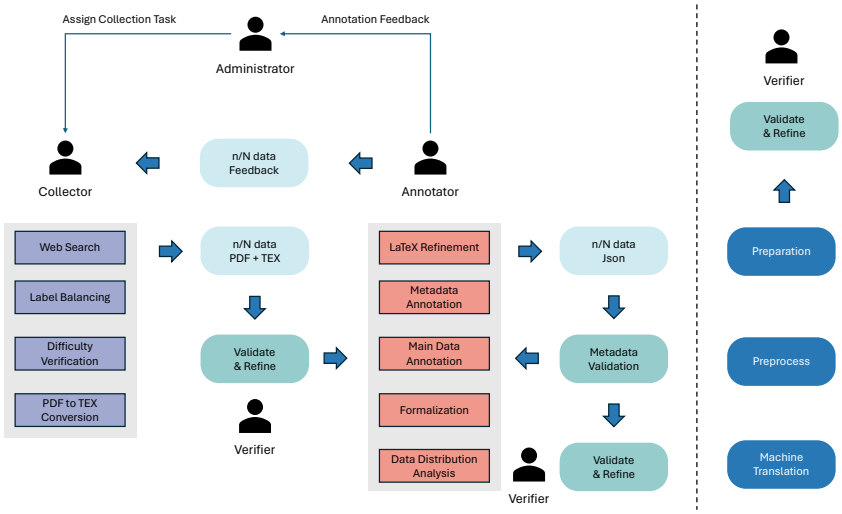


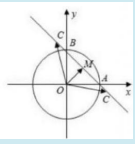
Figure 7: Pipeline invloving data collection, annotation and verification.

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

ID: 719

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

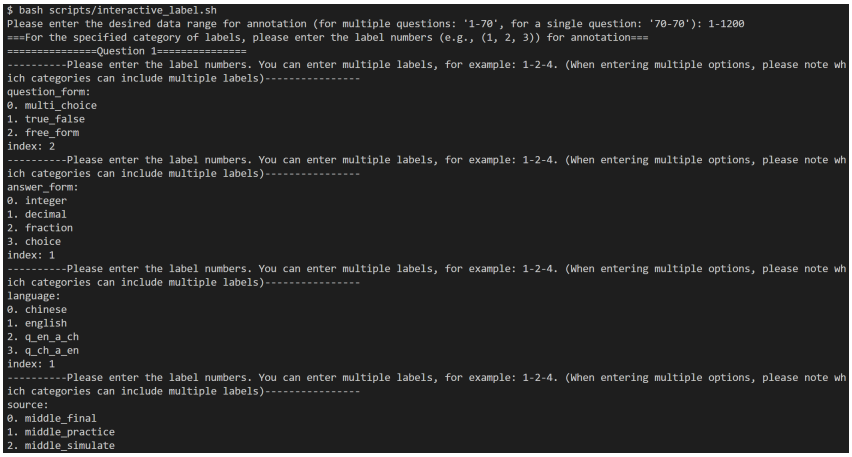
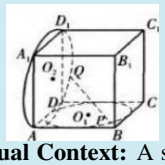


Figure 8: Annotation tool for interactive labeling

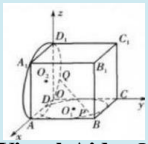
Table 12: Example of mathematical problem with rectangular three-dimensional coordinate system as visual aid.

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 .

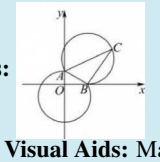
Visual Aids: 

Caption of Visual Aids: Let the origin Oand Dpoints coincide, along DD_1 Establish the Zaxis in the DCdirection, establish the yaxis in the DAdirection, and establish the Xaxis in the DCdirection

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

ID: 723

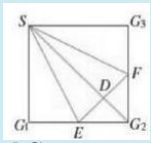
Question: Given that the plane vector $\vec{a}, \vec{b}, \vec{c}$ satisfies the angle between $|\vec{a}| = 1, |\vec{b}| = \sqrt{3}, \vec{a} \cdot \vec{b} = 0, \vec{c} - \vec{a}$ and $\vec{c} - \vec{b}$ is $\frac{\pi}{6}$, then the maximum value of $\vec{c} \cdot (\vec{b} - \vec{a})$ is

Visual Aids: 

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

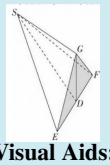
Table 14: Example of mathematical problem with solid geometry as visual aid.

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?

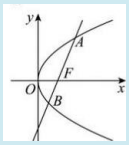
Visual Aids: 

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 15: Example of mathematical problem with function graph 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 Cintersect at two points A, B, If $\overrightarrow{AF} = \lambda\overrightarrow{FB}$, $\lambda = (\quad)$

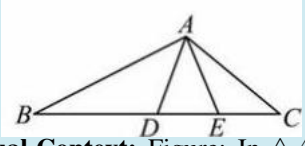
Visual Aids: 

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 Cintersect at two points A, B

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

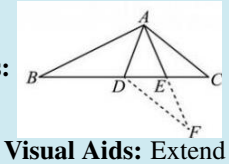
ID: 30

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: $AB = 3AE$.

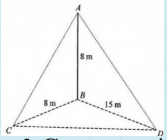
Visual Aids: 

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

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

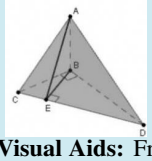
ID: 650

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 θ, then $\tan\theta =$

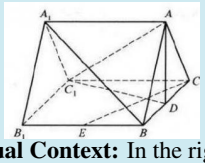
Visual Aids: 

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

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

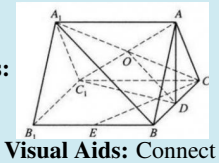
ID: 1024

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: 

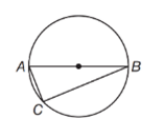
Caption of Visual Aids: Connect A_1C , intersect with AC_1 at O point, connect OD

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

Category	**Detail**
Auxiliary Line	1.1 Connect the points a,b to make the line segment 1.2 Connect points a,b and extend them to intersect the CD line at point E. 1.3 Make a vertical line AB through point a.
Rectangular Three-Dimensional Coordinate System	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.
Rectangular Coordinate System	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 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
Function Graph	4.1 Draw the $y^2 = ax$ image in the real coordinate system with 0 as the origin. 4.2 Draw the $y^2 = ax$ image in the real coordinate system with 0 as the origin and the y=ax image intersecting at point k in the first quadrant and at point D in the second quadrant.
Riemann integral problem	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.

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

Visual Context:

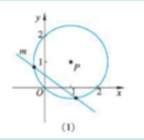


Question: $\overline{AB}$ is a diameter, $AC = 8$ inches, and $BC = 15$ inches. Find the radius of the circle.

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: AB2=AC2+BC2 Substitute the given values: AB2=82+152=64+225=289 Thus, AB=289=17 Since AB is the diameter, the radius r is half of AB: r=172=8.5 The correct answer is (C) 8.5.

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

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.

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:

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 ...

F Detail Experiment Settings

F.1 Hyperparameters

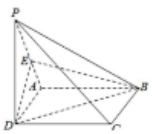
We utilize the default inference settings for each LLMs and LMMs in our experiments. Only specific hyperparameters that are necessary to clarify are listed in Table 23 and 24. We conduct open source model inference based on [37]. We report a single run score for each experiment.

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

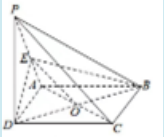
Table 22: A successful Direct Visual-Aided Reasoning example

Visual Context:



Question: Given that $PD \perp$plane $ABCD$, $ABCD$ is A square, and E is the midpoint of PA, judge :(1) $PC//$plane EBD

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:$\because$ EO is parallel to PC (as midlines in a trapezoid).$\because$ PC $\not\subset$ plane EBD and EO $\subset$ plane EBD.$\therefore$ PC is parallel to plane
$mathrmEBD$.So the answer is True.

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

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

instructions for each task are listed in Table 25. ICL examples can be found at our code open sourced after acceptance.

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

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-vl-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

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 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.5 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 More Experimental Results

More experimental results are in Figure 9.

G.1 Results of other tasks

G.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.

Table 25: Instructions for different mathematical problem solving tasks. CQ2A and pCQ2A correspond to General Reasoning. CQ2VA and pCQ2VA correspond to Direct Visual-aided Reasoning. CQpV2A and pCQpV2A correspond to Indirect Visual-aided Reasoning.

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,

G.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. Doubao-Seed-1.6 outperforms other models in all math branches. Notably o4-mini achieves comparable results toward Doubao-Seed-1.6 in "calculus" branch when reason with provided visual aids, exhibiting robustness and enhanced capabilities in visual reasoning under "calculus". GPT-4.1 and InternVL3.5-38B share similar performance in GR and D-VAR task in "analytic geometry" and "calculus". However, performance of InternVL3.5-38B drops significantly, demonstrating higher robustness of GPT-4.1.

The accuracy scores of both LLMs and LMMs 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

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

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 denominator 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'.

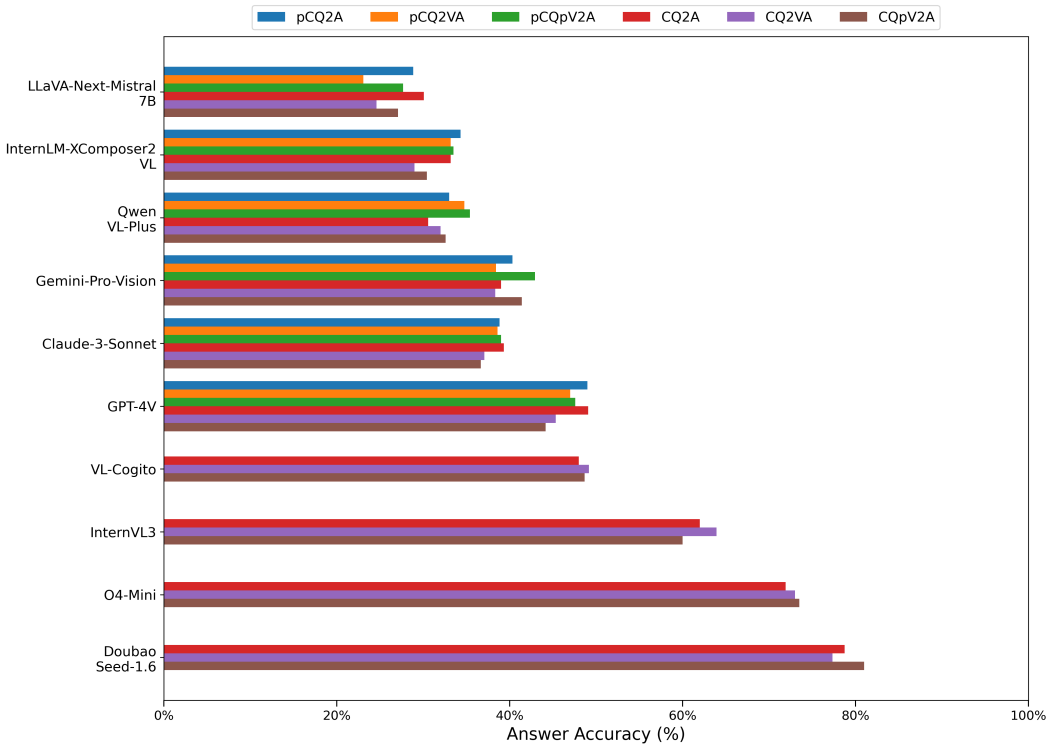


Figure 9: CQ2A is corresponding to GR, CQ2VA is corresponding to D-VAR, and CQpV2A is corresponding to I-VAR.

also attains the best score at "solid geometry" branch, manifesting robust visual context utilization within "solid geometry" branch. GPT4V and GPT4 share 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"

Table 27: 3 ICL examples leveraged to prompt GPT-4 for answer extraction. Full 5 examples will be listed in code open sourced after acceptance.

Element	Prompt
Example 1	**Instruction:** Answer the question requiring a integer answer and provide the final value at the end (e.g., -1, 0, 1, 2, 3). **Question:** Acute Angle $\triangle ABC, BE$vertical ACat E, CDvertical ABat D, given acute Angle $\triangle ABC, BE$vertical ACat E, BCD vertical Ain D, BC $=$ 25, CE $=$ 7, BD $=$ 15, BE, CDto H, connection DE, DEfor circle diameter, and ACto another F, Then the length of AF is **Model response:** Connect DF, then $\angle$DFE $=$ $\angle$BEC $= 90^\circ$, so DF//BE. In Rt $\triangle$BEC, BC $= 25,$CE $= 7$, so BE $= 24, \cos\angle$ACB $= \frac{7}{25}$, $\sin\angle ACB = \frac{24}{25}$, in Rt $\triangle BCD$, $BC = 25, BD = 15$, so $\cos\angle ABC = \frac{3}{5}, \sin\angle ABC = \frac{4}{5}$. So $\sin A = \sin(\angle ABC + \angle ACB) = \sin\angle ABC\cos\angle ACB + \cos\angle ABC\sin\angle ACB = \frac{4}{5}$, in Rt $\triangle ABE$, $\sin A = \frac{BE}{AB} = \frac{24}{AB} = \frac{4}{5}$, we get $AB = 30$, and $BD = 15$, so the point D is the midpoint of AB, and DF is the median of $\triangle ABE$, so $AF = \frac{1}{2}AE = \frac{1}{2}\sqrt{30^2 - 24^2} = 9$. So the final answer is 9 **Extracted answer:** 9
Example 2	**Instruction:** 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) **Question:** In a right angle $\triangle ABC$, $\angle DAC = 45^\circ, AD = 13, AB = 12$. Calculate the length of CD. **Model response:** Let $DE = x$, then $CE = AE = 13 + x$, $\frac{12}{13+x} = \frac{5}{x}$ $x = \frac{65}{7}$, $DE = \frac{65}{7}$ $CE = \frac{156}{7}$, $CD = sqrtDE^3 + CE^2 = \frac{169}{7} = 24.1$ **Extracted answer:** 24.143
Example 3	**Instruction:** 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). **Question:** Linear $x + y = \sqrt{3}$a and round $x^+y^22 = a^2 + (a-1)^2$intersect at point of a,B, $O$is the origin of coordinates, if $triangleisaOB$is triangle, the real value of a **Model response:** From the question, the radius of the circle $r = \sqrt{a^2 + (a-1)^2}$; the distance from the centre of the circle to the line $x + y - \sqrt{3}a = 0$ $d = \frac{\lvert\sqrt{3}a\rvert}{\sqrt{2}}$; because $\triangle AOB$ is a positive triangle, so $d = \frac{\sqrt{3}}{2}r$, i.e. $\frac{\lvert\sqrt{3}a\rvert}{\sqrt{2}} = \frac{\sqrt{3}}{2}\sqrt{a^2 + (a-1)^2}$, which solves $a = \frac{1}{2}$ **Extracted answer:** 1/2

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 geometry" 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

Table 28: 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.

Model	ALL	PLG	SDG	AYG	CAL	AXL	RTC	THC	PLG	SDG	FUG
Heuristics Baselines											
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
Large Language Models (LLMs): Text-Only Input											
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
Large Multimodal Models (LMMs): Text-Only Input											
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
Large Multimodal Models (LMMs): Multimodal Input											
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

of "calculus and functions" can often be solved by generic method, most models maintain the highest accuracy in this branch.

G.2.2 Performance across Complexities

G.2.3 Performance across Visual Aids

With Image as input across different visual aids required to generate, the accuracy scores of mainstream LMMs under three tasks are listed in Figure 22, 23, and 24. Doubao-Seed-1.6 outperforms other models on problems in each category across all three tasks. 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 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.

Table 29: 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.

Model	ALL	PLG	SDG	AYG	CAL	AXL	RTC	THC	PLG	SDG	FUG
Heuristics Baselines											
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
Large Language Models (LLMs): Text-Only Input											
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
Large Multimodal Models (LMMs): Text-Only Input											
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
Large Multimodal Models (LMMs): Multimodal Input											
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

G.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 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** 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 chunck:

$$NG = \{Count(g_1), Count(g_2), ..., Count(g_m)\} \tag{4}$$

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

$$NG_{\text{clip}} = NG_i \cap NG_j \tag{5}$$

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

Then,

$$[h]S = \frac{|NG_{\text{clip}}|}{|NG_{\text{base}}|} \tag{7}$$

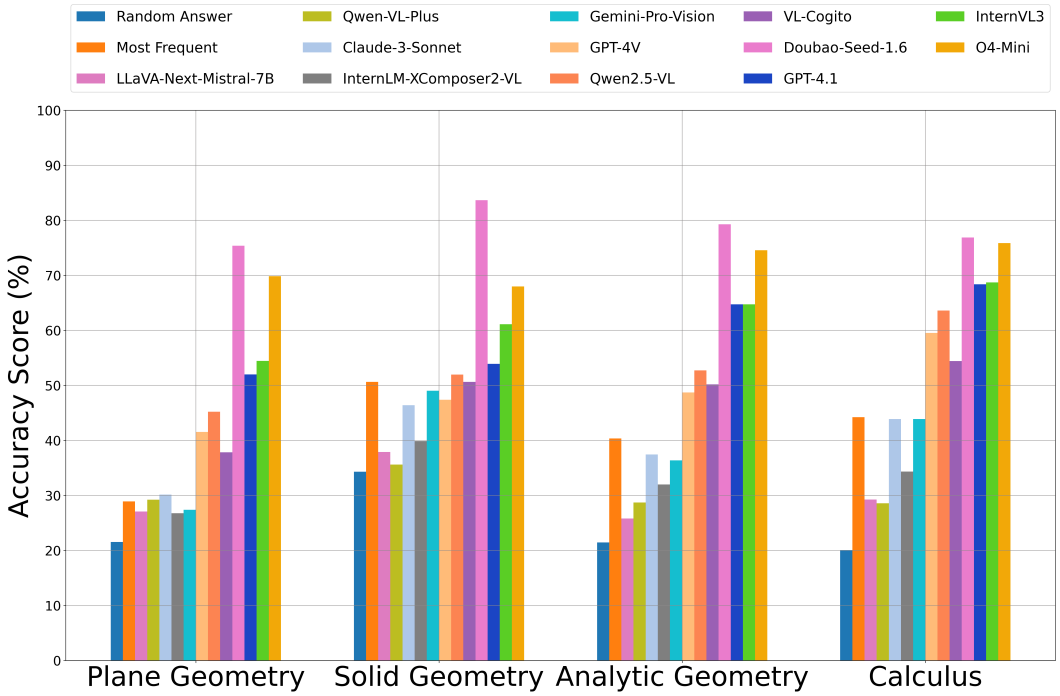


Figure 10: Accuracies of LLMs and LMMs upon CQ2A (General Text-Only Reasoning) task across math branches.

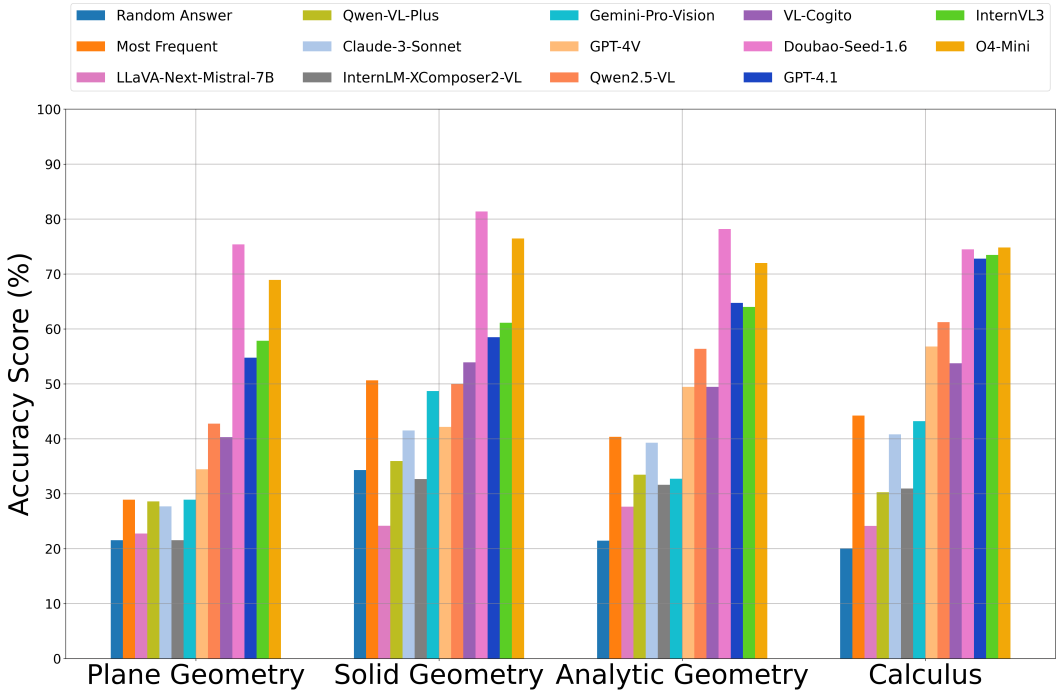


Figure 11: Accuracies of LLMs and LMMs upon CQ2VA (Direct Visual-Aided Reasoning) task across math branches.

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 30. Since models achieve

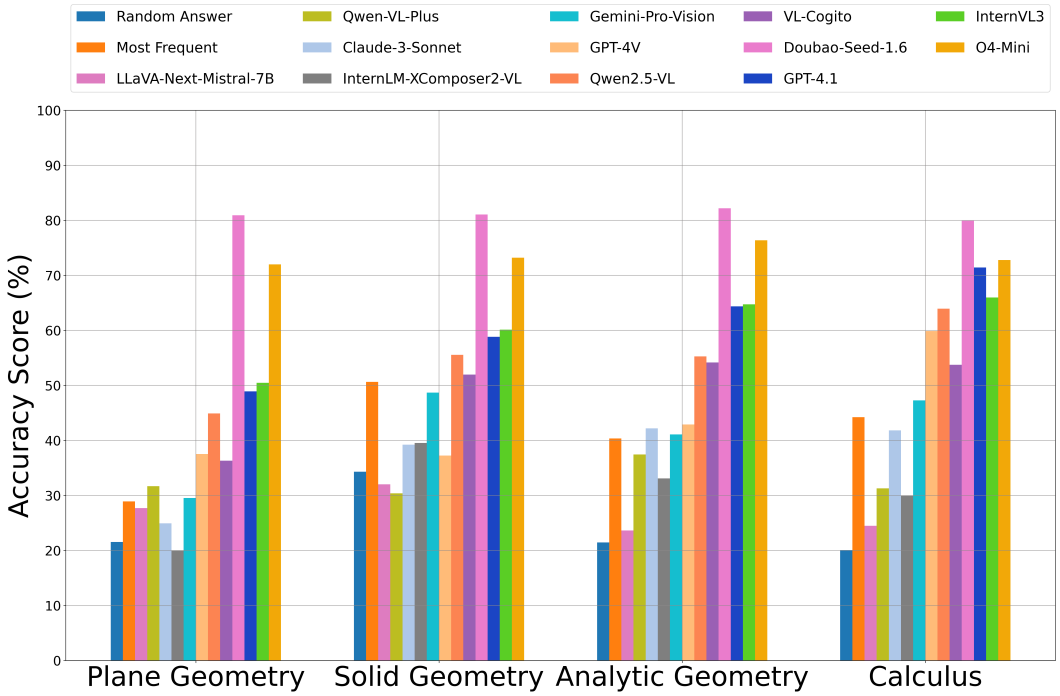


Figure 12: Accuracies of LLMs and LMMs upon CQpV2A (Indirect Reasoning) task across math branches.

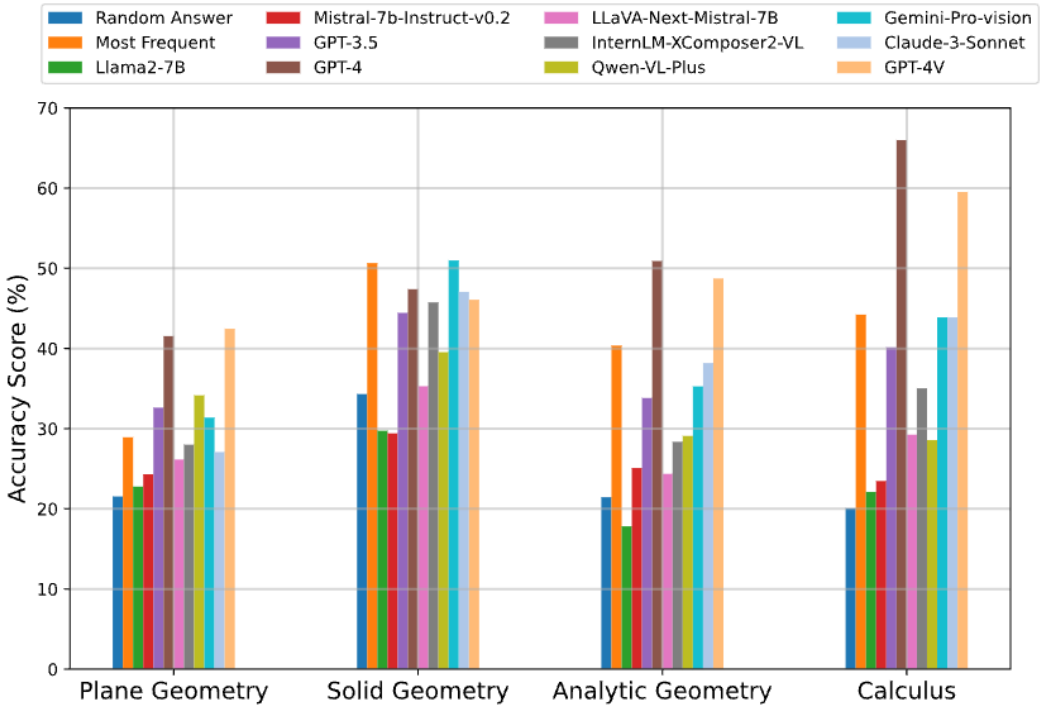


Figure 13: Accuracies of LLMs and LMMs upon pCQ2A (General Text-Only Reasoning) task across math branches.

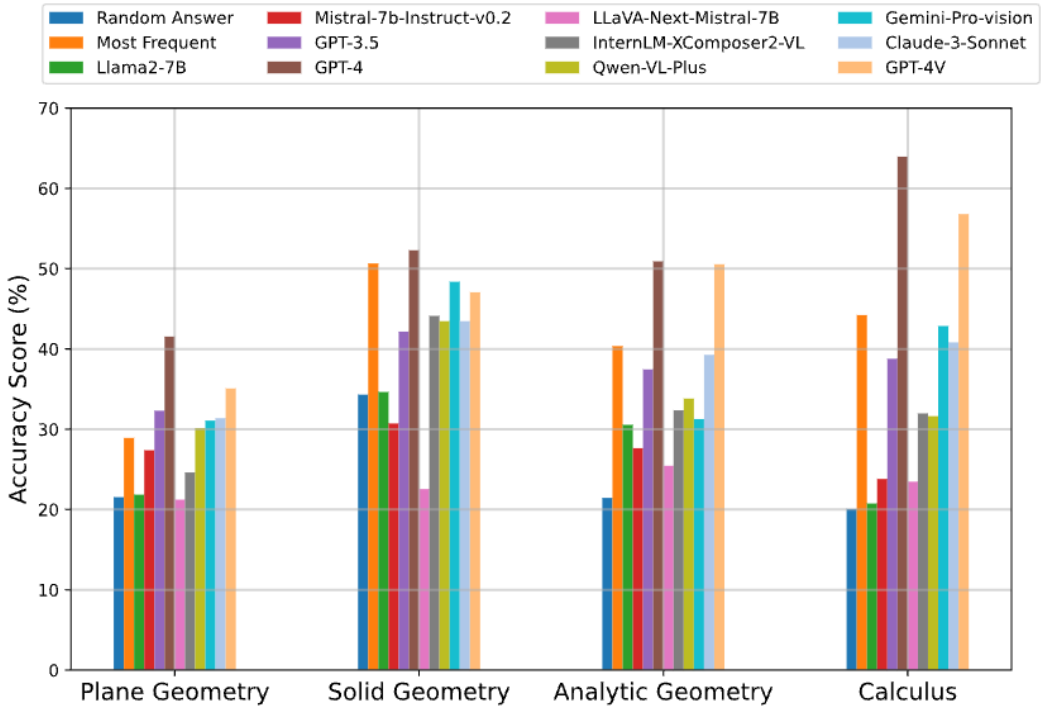


Figure 14: Accuracies of LLMs and LMMs upon pCQ2VA (Direct Visual-Aided Reasoning) task across math branches.

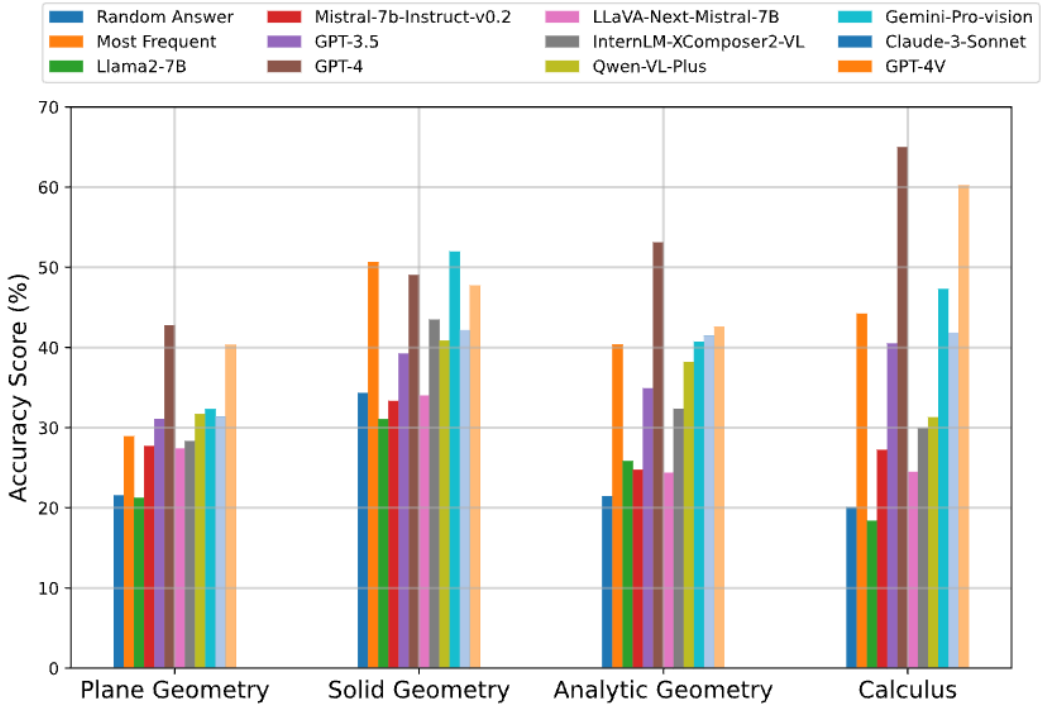


Figure 15: Accuracies of LLMs and LMMs upon pCQpV2A (Indirect Reasoning) task across math branches.

low $S_{dReas-tAid}$ and tend to solve mathematical problems without visual aids, it becomes necessary

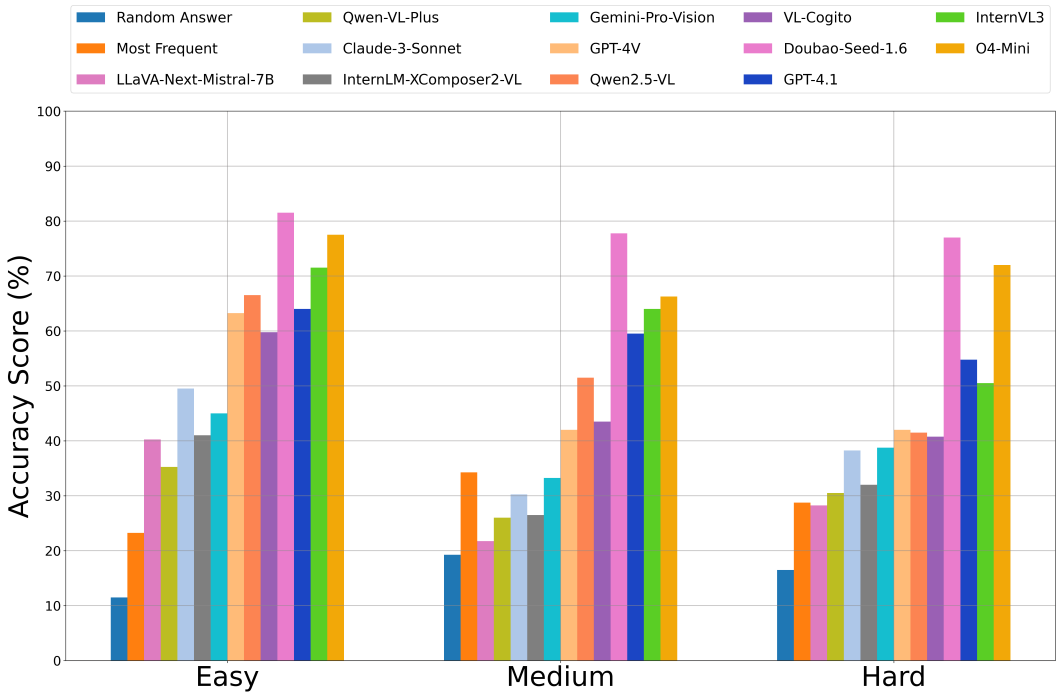


Figure 16: Accuracies of LLMs and LMMs upon CQ2A (General Text-Only Reasoning) task across complexity levels.

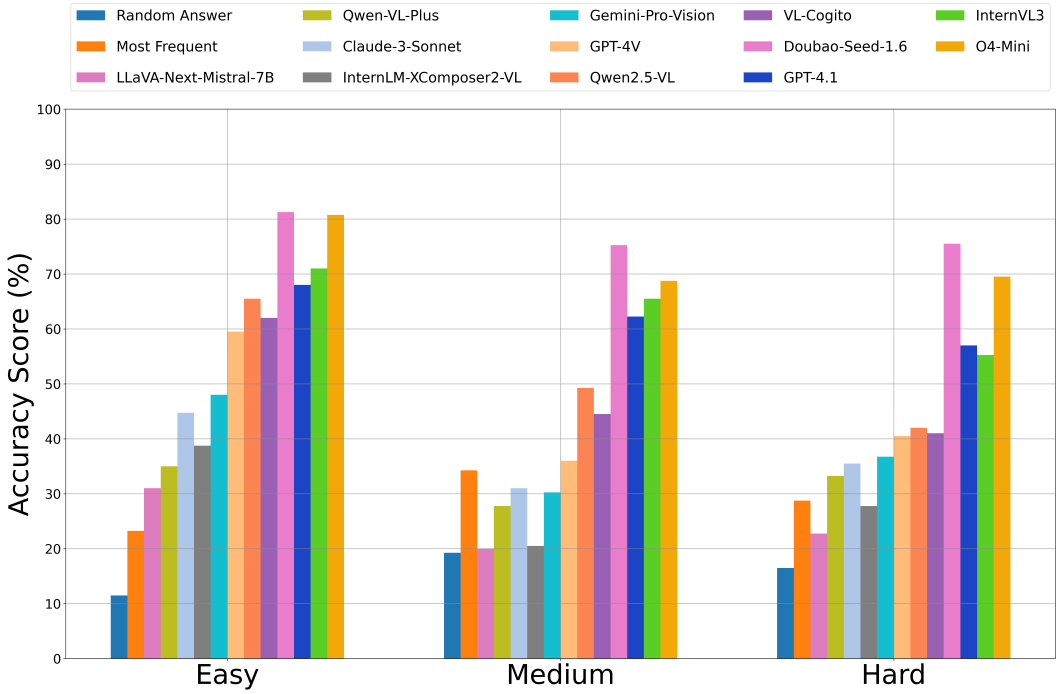


Figure 17: Accuracies of LLMs and LMMs upon CQ2VA (Direct Visual-Aided Reasoning) task across complexity levels.

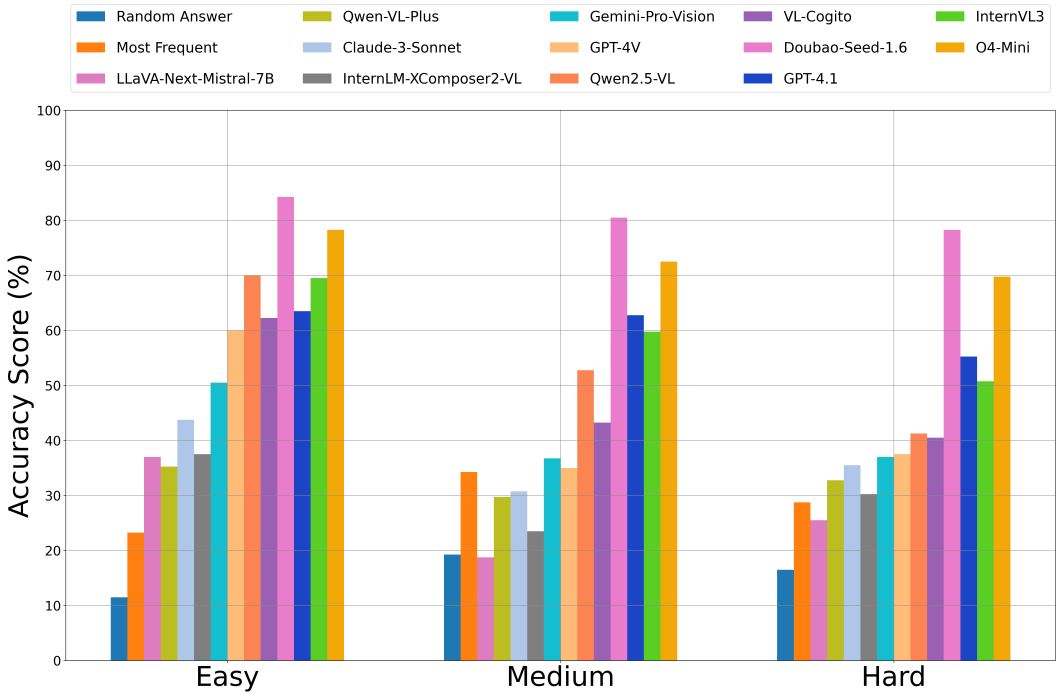


Figure 18: Accuracies of LLMs and LMMs upon CQpV2A (Indirect Reasoning) task across complexity levels.

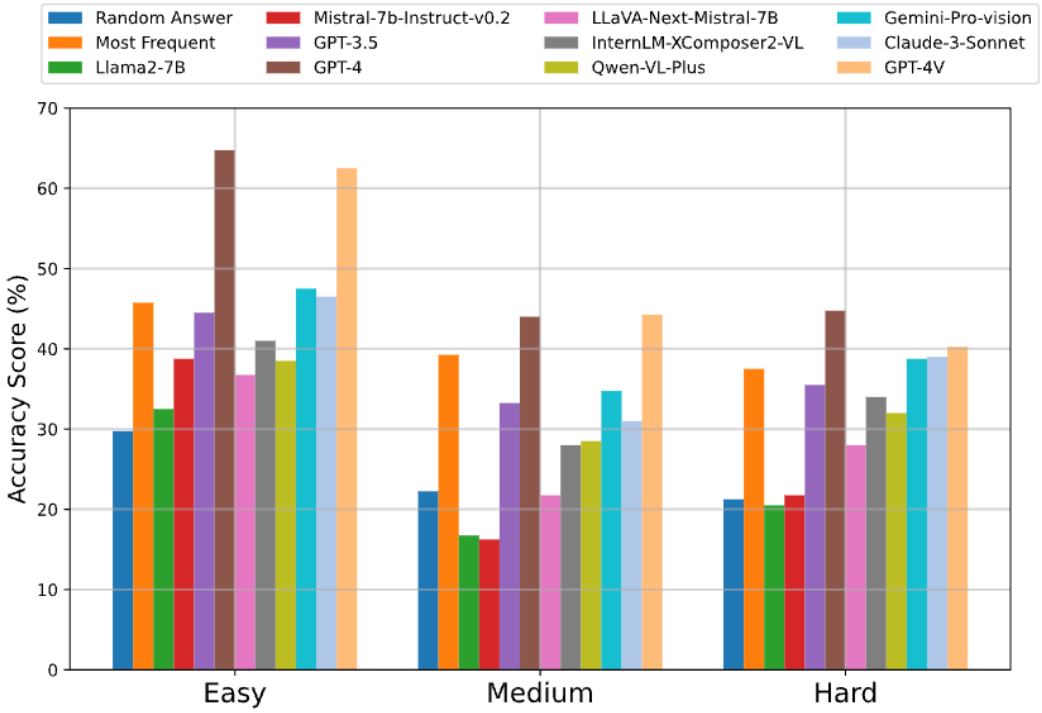


Figure 19: Accuracies of LLMs and LMMs upon pCQ2A (General Text-Only Reasoning) task across complexity levels.

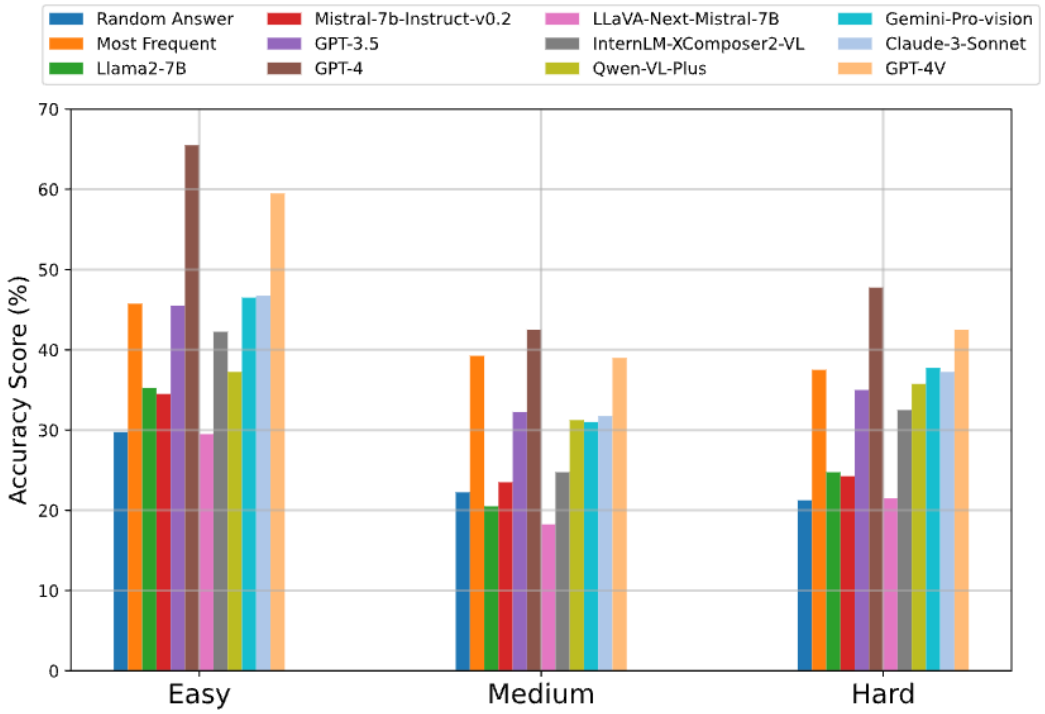


Figure 20: Accuracies of LLMs and LMMs upon pCQ2VA (Direct Visual-Aided Reasoning) task across complexity levels.

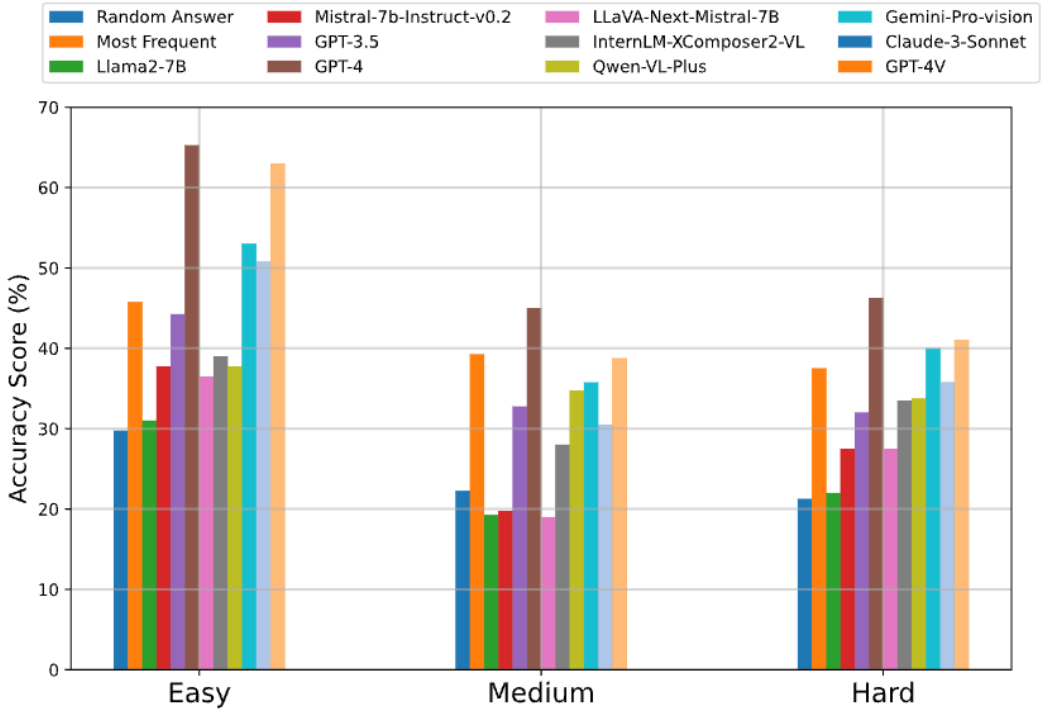


Figure 21: Accuracies of LLMs and LMMs upon pCQpV2A (Indirect Reasoning) task across complexity levels.

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.

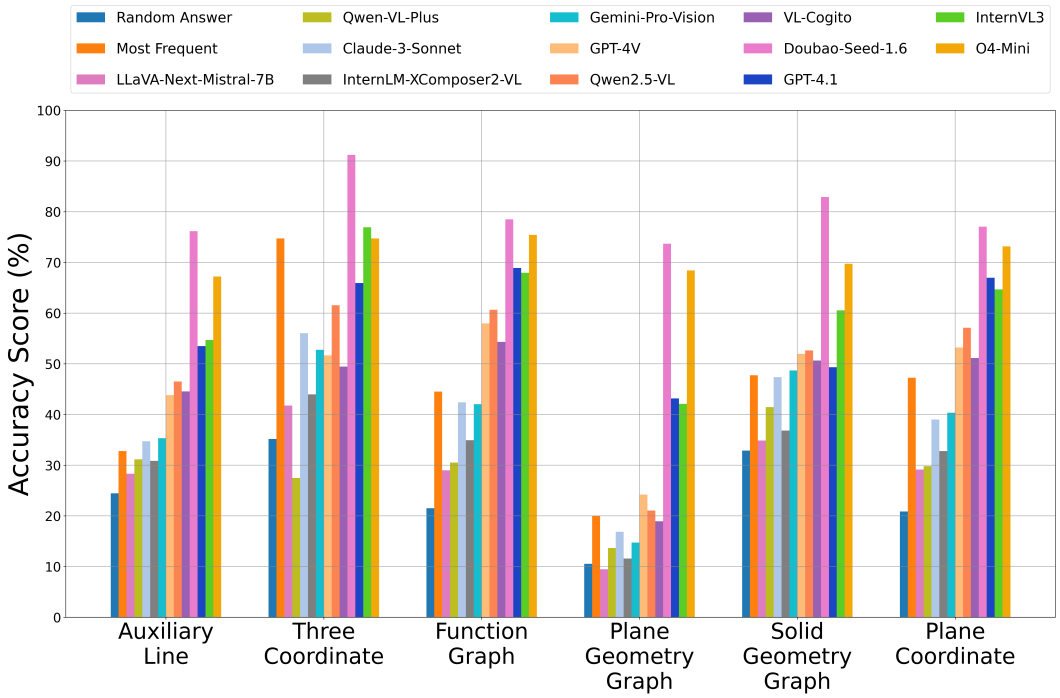


Figure 22: Accuracies of LLMs and LMMs upon CQ2A (General Text-Only Reasoning) task across visual aids.

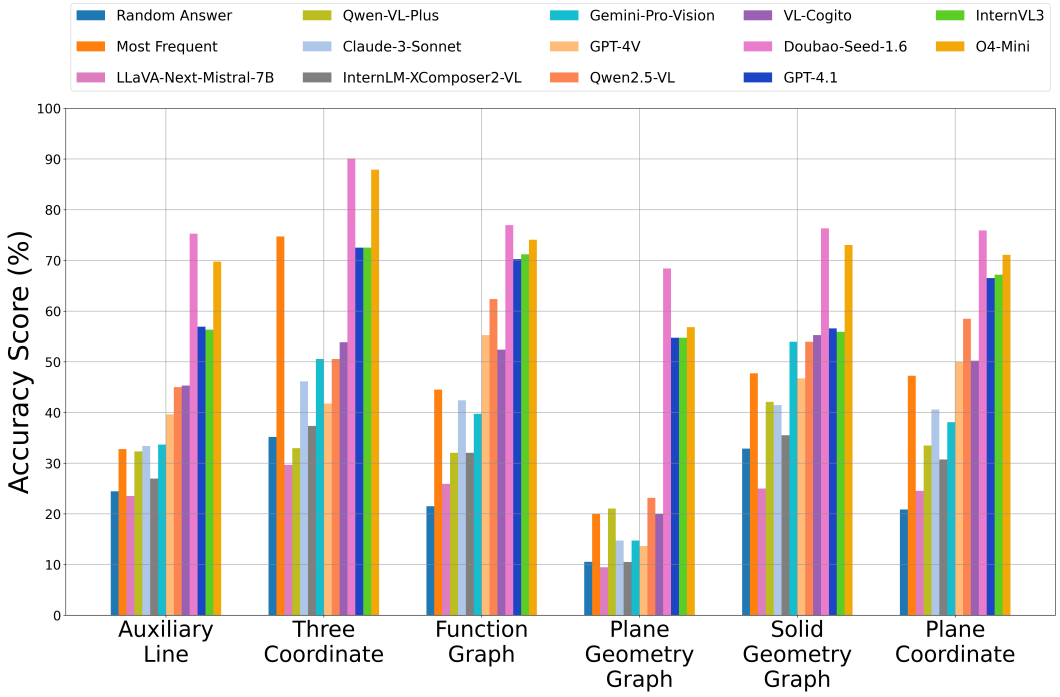


Figure 23: Accuracies of LLMs and LMMs upon CQ2VA (Direct Visual-Aided Reasoning) task across visual aids.

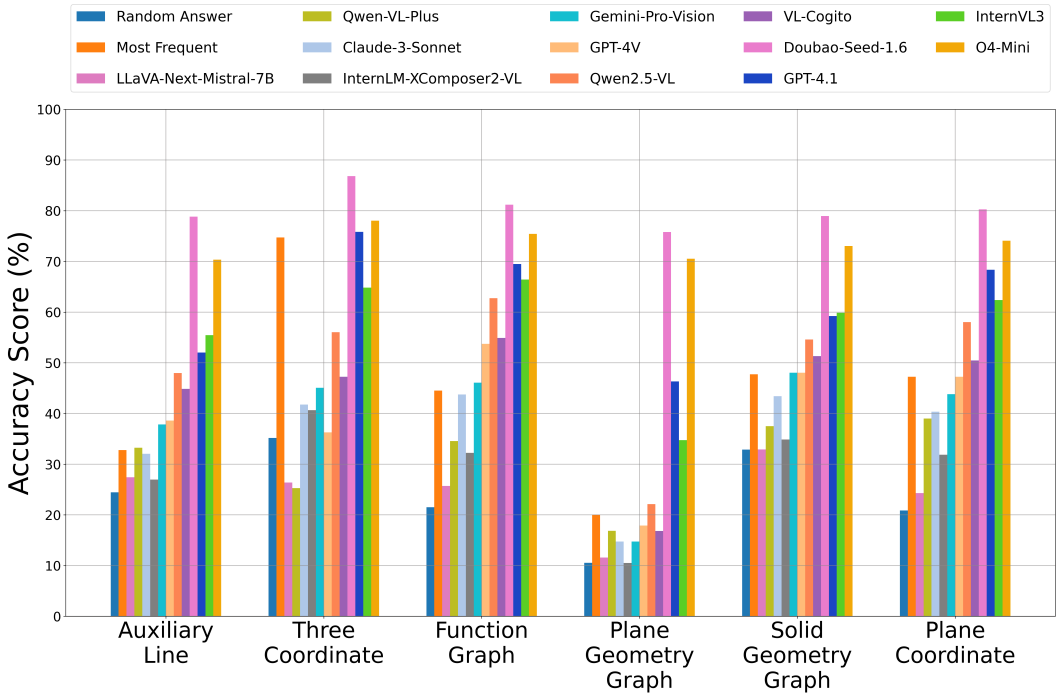


Figure 24: Accuracies of LLMs and LMMs upon CQpV2A (Indirect Reasoning) task across visual aids.

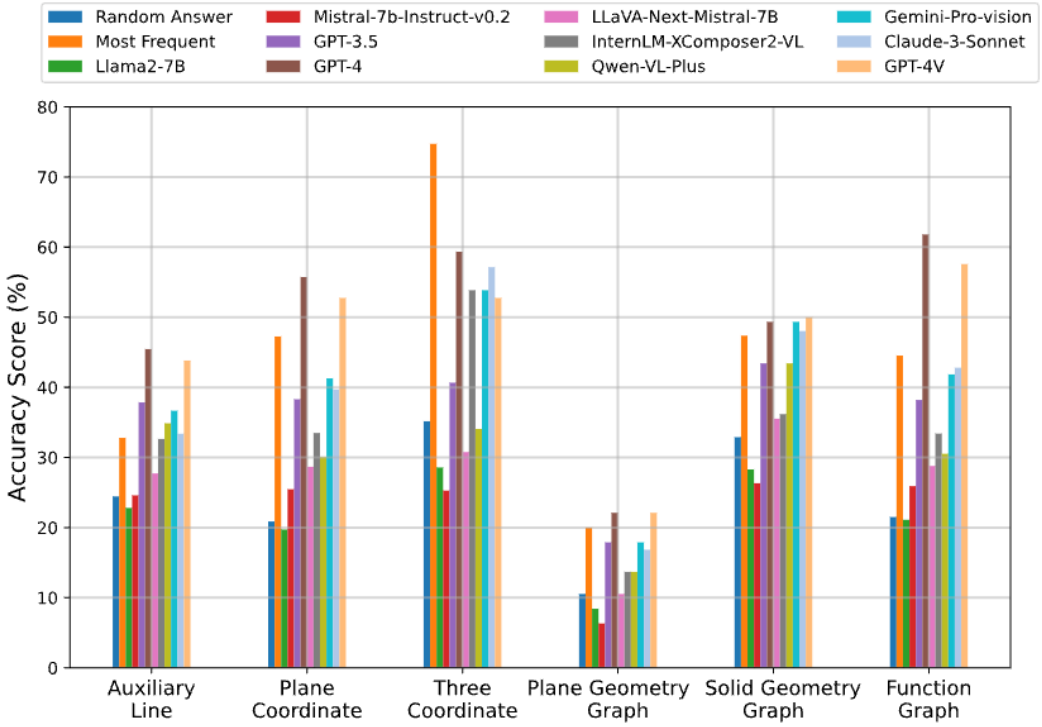


Figure 25: Accuracies of LLMs and LMMs upon pCQ2A (General Text-Only Reasoning) task across visual aids.

H In-depth Model Analysis

H.1 General Reasoning Tendency

H.2 Visual Aid Inference Capability

As shown in Tables 37 and 38, GPT4V demonstrates varying performance in generating visual aids, both correctly and incorrectly.

I 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 35.

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.

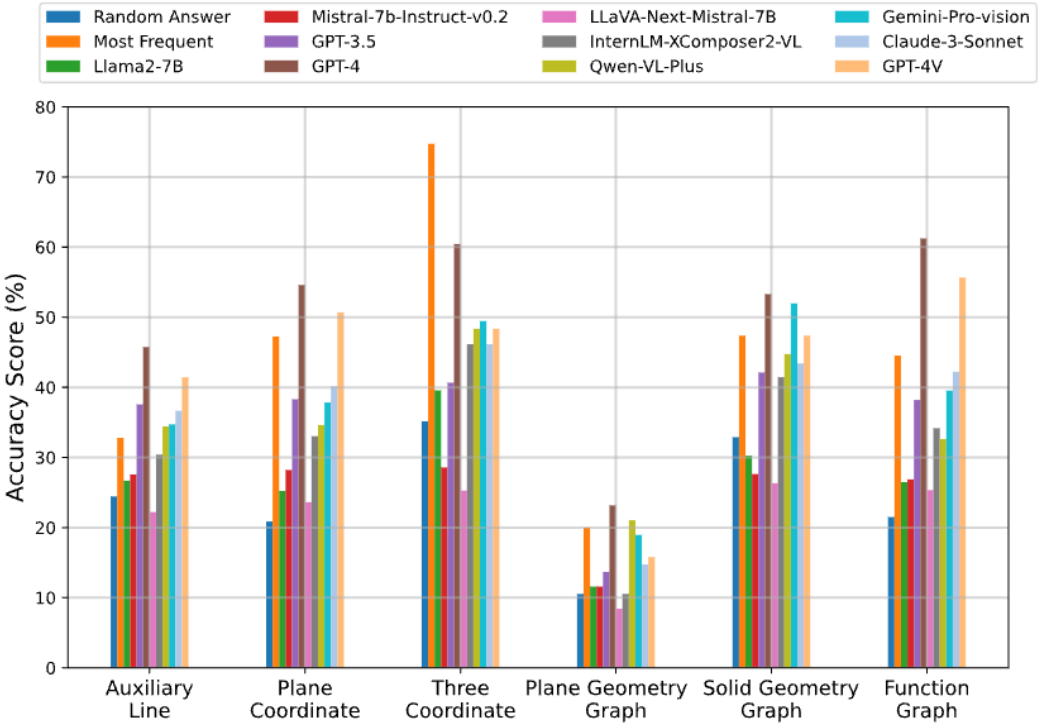
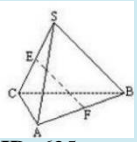


Figure 26: Accuracies of LLMs and LMMs upon pCQ2VA (Direct Visual-Aided Reasoning) task across visual aids.

Table 30: Three types of similarity scores reported across different modles and modalities.

Model	$S_{hAid-dReas}$	$S_{hAid-tAid}$	$S_{dReas-tAid}$
Large Language Models (LLMs): Text-Only Input			
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
Large Multimodal Models (LMMs): Text-Only Input			
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
Large Multimodal Models (LMMs): Multimodal Input			
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 31: Example of GPT4V solve mathematical problem by performing text-only reasoning to achieve correct result.



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° B. 60° C. 45° D. 30°

Table 32: 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°. So, the final answer is C. 45°..

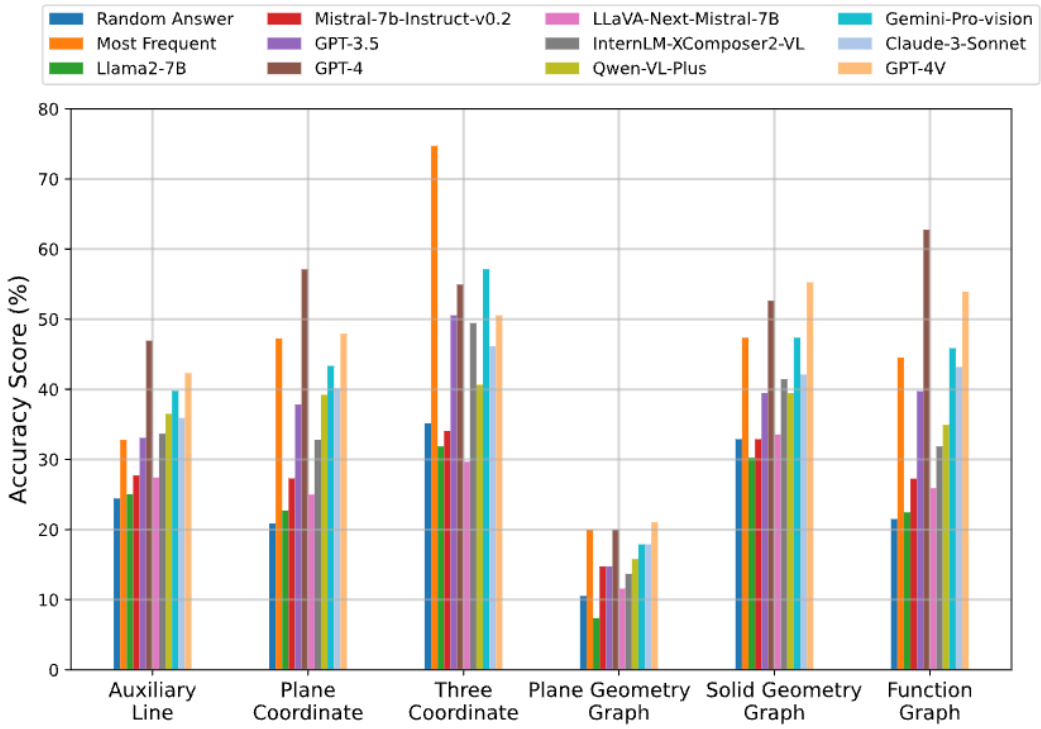


Figure 27: Accuracies of LLMs and LMMs upon pCQpV2A (Indirect Reasoning) task across visual aids.

Table 33: Example of GPT4V solve mathematical problem by backward reasoning from potential solutions.

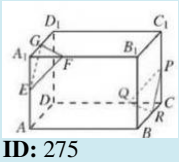
ID: 466

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 34: 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

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 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.

J 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 [38] and MPS [39] have been extensively explored. On the multi-modal side, [40, 41] 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 [6]. More recently, [7] established a comprehensive benchmark that incorporates different visual contexts.

Table 35: Example of GPT4V solve mathematical problem by conducting arithmetic calculation.

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:

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

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

$$AB = \sqrt{4cdot14 + 4cdot14}$$

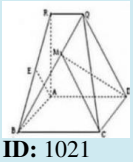
$$AB = \sqrt{8cdot14}$$

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

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

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.

Table 36: Example of GPT4V provide correct answer due to hallucination.



ID: 1021

Question: As shown in the figure, $RA \perp$ is the plane $ABCD, RQ//AD$, 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 $RQ//AD$, 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.

LLMs for Mathematical Reasoning LLMs have not achieved satisfactory performance in mathematical domain under zero or few-shot settings [42]. Chain-of-thought reasoning and prompt engineering methods are introduced to improve step-wise reasoning and decoding control [43]. In multi-modal setting, LLMs can leverage rich information from visual contexts for subsequent mathematical reasoning. [32, 24] explore reasoning over diverse figures that contain abundant numerical and spatial information. Interaction with external tools [44] and downstream instruction tuning [45] are also widely employed to improve overall reasoning quality. A relevant idea proposed by [25] explores inter-connected 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 [46, 47]. Notably, works like [15, 48] 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

Table 37: Example of GPT4V generate visual aids correct in semantics.

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. $3\int_0^{\pi} \cos\theta d\theta$

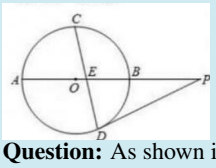
Ground Truth Visual Aids: Draw images of polar curves $r = 2\cos\theta$ and $r = \cos\theta$

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

Table 38: Example of GPT4V fail to generate visual aids due to task misunderstanding. Model replicate.

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

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 A B as the origin, the line where A B is located as the x axis, and the vertical bisector of line segment A B 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.

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.

K limitation and social impact

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.

L LLM Usage Declaration

We utilized a large language model (LLM) to assist in the writing and editing process of this manuscript. The LLM's role was strictly limited to improving grammar, refining phrasing, and enhancing readability. The core research ideas, experimental design, data analysis, and scientific conclusions were conceived and executed entirely by the authors.

Table 40: Example of GPT4V fail to correctly generate visual aids due to fact 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 $\overline{PQ} \| \overline{CA}$ and $\overline{PR} \| \overline{CB}, \angle CAB = \angle PQR$ 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 41: Example of GPT4V fail to correctly generate visual aids due to context conflicting hallucination

ID: 452

Question: It is known that G is the angle 0 of the triangle ABC . A line through G intersects the line AB, AC at two points M, N , $\overrightarrow{AB} = m\overrightarrow{AM}, \overrightarrow{AC} = n\overrightarrow{AN}$, ($m$. n All are positive numbers), the minimum value of $\frac{1}{m} + \frac{2}{n}$ is ( )

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

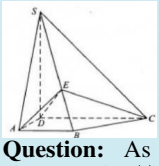
Ground Truth Visual Aids: Draw a triangle $ABC$and draw MGN according to the meaning of the question. Let Dbe the midpoint of BC

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 $\overrightarrow{AM}$ and $\overrightarrow{AN}$. To visualize the relationships between m and n, and their effect on the line ratios, incorporate the vector expressions $\overrightarrow{AB} = m\overrightarrow{AM}$ and $\overrightarrow{AC} = n\overrightarrow{AN}$. 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 $\overrightarrow{AB} = m\overrightarrow{AM}$, and hence $\overrightarrow{AM} = \frac{1}{m}\overrightarrow{AB}$; similarly, $\overrightarrow{AN} = \frac{1}{n}\overrightarrow{AC}$. With these vectors, we can express the position vectors of M and N as $\overrightarrow{AM} = \frac{\overrightarrow{AB}}{m}$ and $\overrightarrow{AN} = \frac{\overrightarrow{AC}}{n}$, 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 42: Example of GPT4V generate different visual aids for alternative substantial reasoning path and provide correct final result

ID: 1023



Question: As shown in the figure, in the quadrangular pheasant $S-ABCD$, $SD \perp$ the base $ABCD, AB//DC, AD \perp DC, AB = AD = 1$, $DC = SD = 2, E$ 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 $BK \perp EC, whereK$is perpendicular, because plane $EDC \perp$is 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.

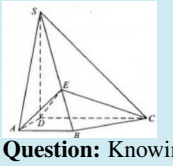
Reasoning: ** By the problem statement, SD is the height of the pyramid, and we know $SD = 2$. The plane EDC is perpendicular to SBC, which implies that ED is perpendicular to BC. Because $ABCD$ is a rectangle with $AB = AD = 1$, we have $BC = DC = 2$. Thus, EDC forms a right-angled triangle with $DC = 2$ 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:

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

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

Table 43: Example of GPT4V generate different visual aids for alternative substantial reasoning path and provide wrong final result

ID: 1060



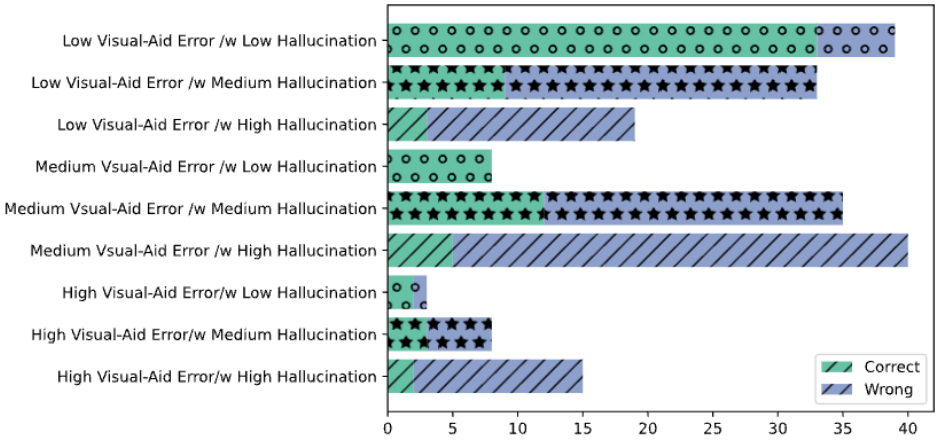
Question: Knowing that there is a point P outside the plane where the convex quadrilateral $ABCD$ is located, and knowing that E, F, G, H, M, N are the midpoints of AB, PC, AD, BC, EF, and GH respectively, then

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 $PD \neq 4MN$

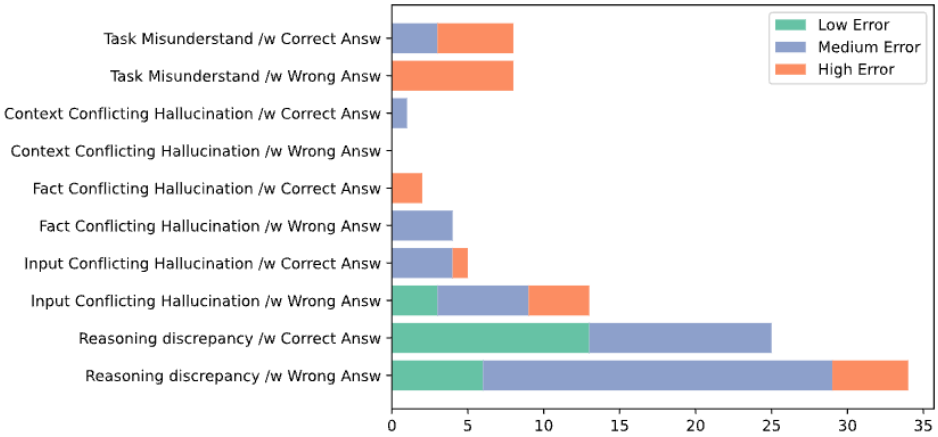
Ground Truth Visual Aids: Taking the midpoint Q of C D and concatenating the midpoints of each side of a quadrilateral A B C D in the same order yields a parallelogram G E H Q and concatenates Q E By dividing the diagonals of a parallelogram equally, we know that QEpasses through the midpoint $N$of GH, and with Nas 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.



(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).