DynaSolidGeo: A Dynamic Benchmark for Genuine Spatial Mathematical Reasoning of VLMs in Solid Geometry

Changti Wu[1,2]*, Shijie Lian[3,2]*, Zihao Liu[4,2], Lei Zhang[1]†, Laurence Tianruo Yang[5,3], Kai Chen[6,2]†
[1]East China Normal University [2]Zhongguancun Academy
[3]Huazhong University of Science and Technology [4]Peking University
[5]Zhengzhou University [6]Zhongguancun Institute of Artificial Intelligence

Abstract

Solid geometry problem solving demands spatial mathematical reasoning that integrates spatial intelligence and symbolic reasoning. However, most existing multimodal mathematical reasoning benchmarks focus primarily on 2D plane geometry, rely on static datasets prone to data contamination and memorization, and evaluate models solely by final answers, overlooking the reasoning process. To address these limitations, we introduce DynaSolidGeo, the first dynamic benchmark for evaluating genuine spatial reasoning in Vision-Language Models (VLMs). Constructed through a semi-automatic annotation pipeline, DynaSolidGeo contains 503 expert-curated seed questions that can, in principle, dynamically generate an unbounded number of diverse multimodal text-visual instances. Beyond answer accuracy, we incorporate process evaluation based on expert-annotated reasoning chains to measure logical validity and causal coherence. Experiments across representative open-source and closed-source VLMs reveal large performance gaps, severe degradation in dynamic settings, and poor performance on tasks requiring high-level spatial intelligence, such as mental rotation and visualization. The code and dataset are available at DynaSolidGeo.

1. Introduction

Geometry problem solving has long played a central role in mathematical reasoning, requiring integrating visual understanding and symbolic reasoning across complex graphic and textual contexts [38]. According to structural properties, geometry can be categorized into plane geometry and solid geometry. Compared to plane geometry, solid geometry imposes substantially higher spatial mathematical reasoning ability, as reasoning in three dimensions entails spatial intelligence, including spatial perception, spatial relation, spatial orientation, spatial rotation, and spatial visualization that goes beyond two-dimensional recognition [14, 16, 22, 29]. Such tasks remain difficult even for human learners [7], and represent a formidable open challenge for current AI systems.

In parallel, recent years have witnessed remarkable progress in multimodal large language models (MLLMs). Building on the successes of foundation models, vision-language models (VLMs) [4–6, 11, 17, 23, 27] have rapidly advanced the state of the art in a wide spectrum of multimodal understanding tasks. Among these tasks, multimodal mathematical reasoning has emerged as a challenging yet vibrant frontier, with benchmarks such as GeoQA [8], MathVista [25], and GeoSense [38] exposing both the promise and the limitations of current VLMs. These carefully designed benchmarks have played a pivotal role in advancing the field, providing standardized evaluation and catalyzing iterative improvements in model design and training paradigm.

Despite this progress, current multimodal mathematical reasoning benchmarks exhibit critical limitations. First, the vast majority of existing geometry-related benchmarks focus on plane geometry or diagram-based word problems, leaving solid geometry, which places heightened demands for spatial intelligence, largely underexplored. For example, PGPS9K [42] contains more than 9,000 plane geometry questions but no solid geometry items, and in GeoEval [41], tasks involving solid geometry constitute merely 2% of the benchmark. Second, nearly all existing multimodal mathematical reasoning benchmarks are static, relying on fixed and finite test sets that are susceptible to data contamination and memorization. Recent analyses demonstrate that large models can memorize and regurgitate benchmark data [10, 12, 21, 26, 28], and some studies show that decontaminated re-releases often lead to substantial drops in performance [44], confirming that static evaluation may significantly overestimate true reasoning and generalization [31]. Similar concerns have motivated dynamic evaluations in coding [18, 45] and general-purpose

*These authors contributed equally
†Corresponding author

Benchmarks	Language	S.G. Size (Proportion)	S.G. Category	Level	Static/Dynamic	Modality	Metric
GeoQA [8]	EN&CN	0 (0.0%)	-	K.	Static	Text&Image	A.E.
PGPS9K [42]	EN	0 (0.0%)	-	K.	Static	Text&Image	A.E.
MMMU-MATH [40]	EN	0 (0.0%)	-	U.	Static	Text&Image	A.E.
GeoEval [41]	EN	100 (2.0%)	-	K.	Static	Text&Image	A.E.
MATH-Vision [32]	EN	263 (8.7%)	-	K.&U.	Static	Text&Image	A.E.
OlympiadBench [15]	EN&CN	784 (9.2%)	-	C.	Static	Text&Image	A.E.
MathVerse [43]	EN	119 (15.1%)	-	K.	Static	Text&Image	A.E.
GeoSense [38]	EN&CN	350 (20.0%)	2	K.	Static	Text&Image	A.E.&P.E.
SolidGeo [33]	EN&CN	3113 (100.0%)	8	K.&U.&C.	Static	Text&Image	A.E.
GeoLaux [13]	CN	0 (0.0%)	-	K.	Static	Text&Image	A.E.&P.E.
DynaMath [47]	EN&CN	15 seed questions (3.0%)	-	K.&U.	Dynamic	Text&Image	A.E.
DynaSolidGeo (Ours)	EN&CN	503 seed questions (100.0%)	8	K.&C.	Dynamic	Text&Image&Video	A.E.&P.E.

Table 1. Comparison with existing geometry-related mathematical reasoning benchmarks. S.G.=Solid Geometry; **Level**: K.=K-12, U.=University, C.=Competitions; **Metric**: P.E.=Process Evaluation, A.E.=Answer Evaluation.

QA [34, 35], indicating a community-wide shift towards dynamic, contamination-resistant evaluation protocols [9]. Moreover, most existing multimodal mathematical reasoning benchmarks [25, 32, 33, 40–42, 47] evaluate models solely based on answer accuracy, which allows models suffering from data contamination or over-reliance on memorization to appear strong, while failing to reveal their genuine reasoning ability.

To address these limitations, we present DynaSolidGeo, a new benchmark for the dynamic evaluation of VLMs' genuine spatial mathematical reasoning in solid geometry. Unlike existing static resources, DynaSolidGeo consists of 503 seed questions of solid geometry problem solving, each represented by a Python program paired with a corresponding MATLAB program. With the correctness of the question guaranteed, each seed question is parameterized: textual variables in the question statement (e.g., endpoint labels, side lengths, areas, volumes, ratio) as well as rendering parameters of the solid geometry (e.g., camera viewpoints) can all be randomized. By supplying different random seeds, DynaSolidGeo can in principle generate an unbounded number of question-answer instances, where each instance can optionally include two visual versions: a randomized-view image and a 360-degree rotation video. The seed questions of DynaSolidGeo are drawn from diverse and authoritative sources, including China's Gaokao examinations, international mathematics competitions, and widely used training materials for competition preparation. Together, they cover nearly all major categories of high-school and competition-level solid geometry problems (eight in total), including positional relations, angle, length, area, and volume calculations, as well as counting, dynamic, and folding tasks. Moreover, we move beyond answer-only evaluation by incorporating process-level assessment grounded in expert-annotated reasoning chains. Through Answer Accuracy (AA), Process Score (PS), and Process-Qualified Accuracy (PA), we jointly measure answer correctness, reasoning quality, and reasoning-qualified accuracy, offering a more faithful reflection of VLMs' genuine spatial mathematical reasoning ability. To ensure reliability, all solutions are expert-annotated by undergraduates and graduate students from the School of Mathematical Sciences, Peking University, including Chinese Mathematical Olympiad (CMO) gold medalists. A comparative summary with related benchmarks is provided in Table 1.

We evaluate a range of mainstream, latest closed- and open-source VLMs on DynaSolidGeo. Experiments reveal a clear gap between most open-source and closed-source VLMs. Notably, nearly all models struggle with Counting problems, highlighting the lack of higher-order spatial intelligence, such as mental rotation and spatial visualization. Compared to the static source-question dataset, models exhibit a significant performance drop on DynaSolidGeo (up to 20.4% for Claude-Sonnet-4.5), exposing potential data contamination and memorization effects. Furthermore, the additional metric degradation after introducing process evaluation indicates that previous static, answer-only benchmarks likely overestimated model capabilities, whereas DynaSolidGeo provides a more faithful and comprehensive evaluation of genuine spatial mathematical reasoning ability. In summary, our contributions are as follows:

- We design a semi-automatic data annotation pipeline for the seed question annotation of solid geometry problems, which minimizes human involvement without compromising annotation correctness or usability.
- We propose DynaSolidGeo, the first dynamic benchmark for solid geometry problem solving, consisting of 503 carefully curated seed questions that can, in principle, automatically generate an unbounded number of diverse question instances across multiple geometry categories.
- We introduce a process evaluation using expert-annotated reasoning chains that, together with answer evaluation,

provides a holistic measure of VLMs' genuine spatial mathematical reasoning capability.
- We evaluate a series of popular and SOTA VLMs on DynaSolidGeo to gain deeper insights into their spatial mathematical reasoning abilities and conduct extensive analyses, including revealing potential data contamination and memorization phenomenon on static datasets.

2. Related Work

2.1. Multimodal Mathematical Reasoning Benchmarks.

Recent years have witnessed the emergence of multimodal benchmarks that evaluate mathematical reasoning in visually grounded settings. Early efforts include TQA [19] and Geometry3K [24] introduced multimodal reasoning tasks involving diagram-based science and geometry word problems with accompanying 2D visuals. More recent benchmarks, such as GeoQA [8], PGPS9K [42], MMMU-MATH [40], GeoEval [41], MATH-Vision [32], OlympiadBench [15], MathVerse [43], GeoSense [38], and GeoLaux [13] have broadened coverage to thousands of multimodal math problems. However, the vast majority of these resources concentrate on plane geometry and 2D diagrammatic reasoning, leaving solid geometry largely underexplored. A few datasets have attempted to move toward 3D: SolidGeo [33] explicitly targets solid geometry but remains static datasets vulnerable to contamination and memorization; DynaMath [47] introduces dynamic instance generation, but solid geometry is barely represented, with only 15 problems (3%) in the dataset. In contrast, DynaSolidGeo fills this gap with scalable and dynamic solid-geometry coverage.

2.2. Vision-Language Models.

Recent vision–language models (VLMs) such as BLIP-2 [20], Flamingo [3], and LLaVA [23] combine pretrained large language models with visual encoders, enabling open-ended multimodal reasoning and instruction following. Building on this paradigm, the latest generation of VLMs has rapidly advanced in scale, architecture, and reasoning capability. The closed-source models include GPT-5 family [27], Gemini-2.5 family [11], and Claude-Sonnet-4.5 [5], which feature deeply integrated multimodal backbones and enhanced reasoning modules. In parallel, the open-source community has introduced competitive alternatives such as LLaVA-OneVision-1.5 family [4], GLM-4.1V-9B-Thinking [17], Llama-4-Maverick-17B-Instruct [2], InternVL3.5-8B [37], DeepSeek-VL2 [36], and the Qwen3-VL family [30], which push the frontier of visual capabilities. Yet their spatial mathematical reasoning ability remains underexplored, motivating our evaluation on DynaSolidGeo.

3. DynaSolidGeo

We propose DynaSolidGeo, a dynamic multimodal benchmark for spatial mathematical reasoning in solid geometry, which consists of 503 expert-annotated seed questions that can expand into unbounded question–answer instances with randomized text, images, and 360-degree rotation videos, by inputting a random seed.

3.1. Data Collection

The seed questions of DynaSolidGeo are drawn from diverse and authoritative sources to ensure both breadth and rigor. Specifically, we collect 503 solid geometry questions (referred to as source questions) from three major categories: 1) China's Gaokao examinations from 2014 to 2025 (11 years), 2) international mathematics competitions such as the American Invitational Mathematics Examination (AIME), the American Mathematics Competitions (AMC), and the American High School Mathematics Examination (AHSME), and 3) high-level preparation and training materials, including competition handbooks and advanced supplementary textbooks. These sources cover nearly the full spectrum of high-school and competition-level solid geometry categories, encompassing positional relationships, angles, distances, area and volume computation, as well as combinatorial counting, dynamic scenarios, and folding/unfolding problems (see Table 2).

3.2. Data Annotation Pipeline

We design a semi-automatic seed-question annotation pipeline that aims to minimize manual labeling costs while ensuring the correctness and availability of the generated programs. Compared with a fully manual annotation process, this approach substantially reduces human effort. At the same time, in contrast to a fully automatic procedure, it preserves accuracy and reliability in handling complex solid geometry questions. As shown in Figure 1, our data annotation process is divided into the following components:

1. **Expert-Guided Parametrization and Visualization**: Here we follows a *human-in-the-loop* strategy, where human experts collaborate with large models to create a JSON annotation and a MATLAB program for each source question:
   - *JSON-Based Question Parametrization*: For each collected source question, mathematics experts parameterize the question statement by converting fixed values into variable parameters (e.g., endpoint labels, side lengths, areas, volumes, ratios) using f-string syntax, while ensuring correctness and availability. Corresponding answer is also expressed in terms of these variable parameters. An example of parameterized variables is highlighted in red in Fig. 1. Additional metadata (e.g., category and difficulty level) is also included and stored in JSON format.

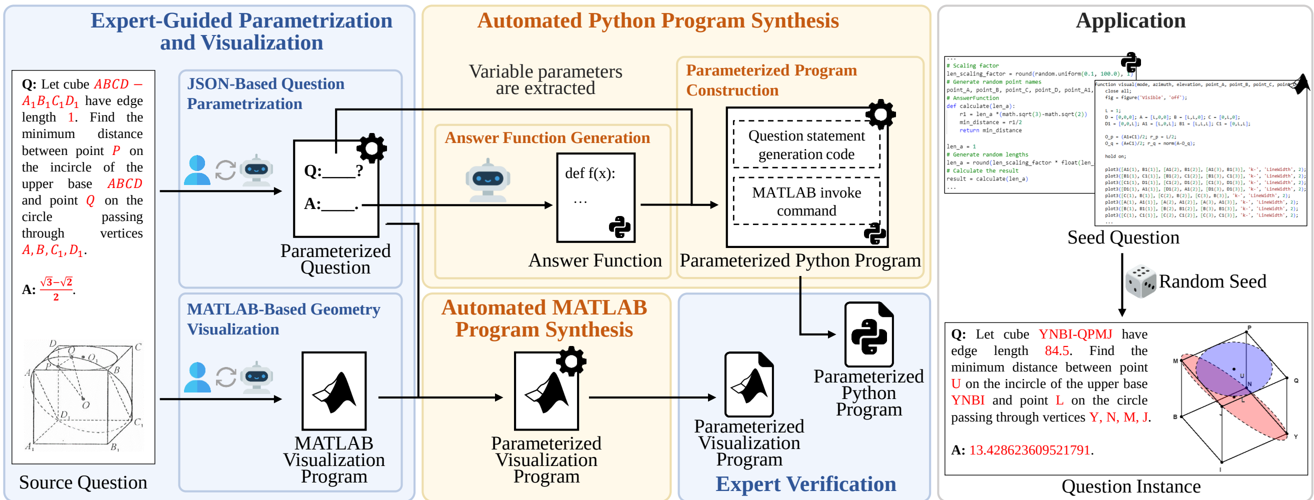

Figure 1. Overview of the data annotation pipeline and the application of seed questions. Annotation: 1) Expert-Guided Parametrization and Visualization: Each source question is first parameterized into a JSON annotation and paired with a MATLAB visualization program. 2) Automated Python Program Synthesis: The pipeline then synthesizes parameterized Python programs that generate textual descriptions and MATLAB invoke commands. 3) Automated MATLAB Program Synthesis: Correspondingly, the pipeline then synthesizes the parameterized version of MATLAB programs for figure and video rendering. 4) Expert Verification: Final human checks ensure the correctness and usability of seed questions. Application: By inputting a random seed, each seed question is instantiated into a question instance.

- *MATLAB-Based Geometry Visualization*: MATLAB experts implement programs that render each solid geometry image and video for each source question.

2. **Automated Python Program Synthesis**:
   - *Answer Function Generation*: With the assistance of the large language model, parametrized answers in JSON are converted into Python functions that dynamically compute results.
   - *Parameterized Program Construction*: A rule-based script automatically assembles parameterized Python programs from the parametrized questions and answer functions. By inputting a random seed, the parameterized Python program randomizes both the MATLAB camera parameters (i.e., azimuth and elevation) and variable parameters in the parameterized question, and finally outputs a JSON entry of the instantiated question along with a MATLAB invoke command.
3. **Automated MATLAB Program Synthesis**: Each MATLAB visualization program is automatically converted into a parameterized version by a rule-based script, aligned with the annotated JSON specification. These programs can be directly invoked by the MATLAB commands generated in the previous step, enabling dynamic rendering of figures and videos consistent with the instantiated question parameters.
4. **Expert Verification**: Final human checks ensure correctness, consistency, and usability of seed questions.

Overall, each seed question is associated with a parameterized Python program for generating the textual description and a parameterized MATLAB program for rendering the corresponding figures and videos. By inputting a random seed, each seed question can be instantiated into a concrete question instance.

3.3. Statistics

Table 2 summarizes the detailed statistics of the DynaSolidGeo dataset. In total, the benchmark contains 503 curated seed questions, all newly constructed for this work. The questions span a diverse set of solid geometry problem categories, including Positional relationship determination (PD, 11.7%), Angle calculation (AN, 20.5%), Length and distance calculation (LC, 13.1%), Area calculation (AR, 11.3%), Volume calculation (VC, 10.3%), Counting problems (CP, 7.0%), Dynamic or moving-point problems (DM, 13.1%), and Folding and unfolding problems (FP, 12.9%). The distribution across difficulty levels is reasonably balanced, with 27.2% easy, 57.7% medium, and 15.1% hard questions. Regarding question types, previous benchmarks (e.g., DynaMath[47], SolidGeo [33], GeoSense[38], MathVerse[43]) generally include multiple-choice questions. Such options inevitably provide strong hints to the models, thereby reducing the difficulty and making it difficult to assess their reasoning ability accurately. In contrast, we rewrite the multiple-choice and proof questions from the source data into fill-in-the-blank formats in our work. As a result, DynaSolidGeo consists of 88.3% numerical questions and 11.7% free-form questions, posing greater challenges to the reasoning ability of VLMs.

Figure 2 shows the distribution of the number of variable parameters contained in the seed questions of DynaSolidGeo. The variable parameters include camera parameters (i.e., azimuth and elevation), endpoint labels, side lengths, areas, volumes, ratios, and so on. As illustrated, the seed questions exhibit substantial variability, highlighting the richness and flexibility of our benchmark design.

Statistic	Number
Total seed questions	503
- Newly curated questions	503 (100.0%)
Categories	
- Positional relationship determination (PD)	59 (11.7%)
- Angle calculation (AN)	103 (20.5%)
- Length and distance calculation (LC)	66 (13.1%)
- Area calculation (AR)	57 (11.3%)
- Volume calculation (VC)	52 (10.3%)
- Counting problems (CP)	35 (7.0%)
- Dynamic or moving-point problems (DM)	66 (13.1%)
- Folding and unfolding problems (FP)	65 (12.9%)
Levels	
- Easy	137 (27.2%)
- Medium	290 (57.7%)
- Hard	76 (15.1%)
Question types	
- Numerical questions	444 (88.3%)
- Free-form questions	59 (11.7%)

Table 2. Statistics of DynaSolidGeo

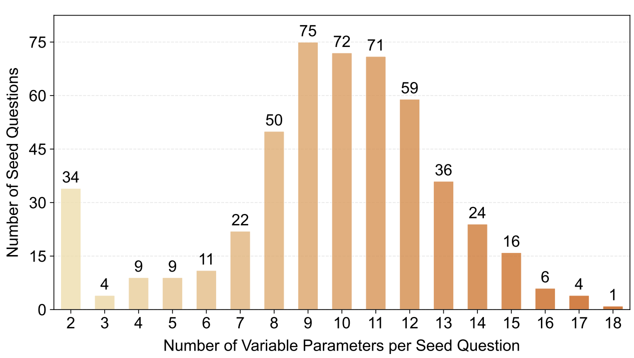

Figure 2. Distribution of variable parameters per seed question.

3.4. Evaluation Metrics

Unlike most existing studies that only assess the correctness of final answers, we additionally introduce process evaluation to holistically assess the model's genuine spatial reasoning capacity. DynaSolidGeo consists of $N = 503$ seed questions, from which we instantiate K batches of question instances by sampling with different random seeds. Building on these instances, we design the following evaluation metrics to systematically evaluate model performance.

3.4.1. Answer Accuracy

We use the Answer Accuracy (AA) to measure the correctness of the model's answers, i.e., the proportion of final answers that are correct on average:

$$\mathcal{AA} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\mathrm{Ans}_{k,i} = \mathrm{GT}_{k,i}], \tag{1}$$

where $\mathrm{Ans}_{k,i}$ and $\mathrm{GT}_{k,i}$ denote the rule-extracted answer and the corresponding ground truth of the i-th question in the k-th batch, respectively.

3.4.2. Process Score

To more accurately assess the quality of the model's reasoning process, we annotate not only parameterized question statements and answers, but also parameterized reference reasoning processes. Each reasoning process is evaluated using LLM as a judge according to the following criteria:

- Logical Alignment: The reasoning presents a coherent derivation whose steps consistently lead to the stated result, with matching variables/units and without any unjustified conclusion jumps.
- No Extraneous Information: The reasoning does not rely on unseen quantities or unsupported facts as essential premises (standard geometric axioms and theorems are allowed).
- Use of Key Dependencies: The reasoning explicitly invokes the key geometric relations specified in the problem (e.g., parallelism, similarity, perpendicularity, collinearity, ratios, or angle constraints), rather than skipping these conditions and merely reporting the final result.

Based on these criteria, we introduce the process evaluation metric, termed the Process Score (PS):

$$\mathcal{PS} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N}\sum_{i=1}^{N}\mathcal{S}_{k,i},$$
$$\text{where} \begin{cases} 0, & \text{if } \mathrm{Ans}_{k,i} \neq \mathrm{GT}_{k,i}; \\ 0.75 \leq \mathcal{S}_{k,i} \leq 1, & \text{if all above criteria are met;} \\ 0 < \mathcal{S}_{k,i} < 0.75, & \text{otherwise.} \end{cases} \tag{2}$$

In Equation 2, $\mathcal{S}_{k,i}$ denotes the process score of the i-th question in the k-th batch scored by the judge model. A higher PS corresponds to a reasoning process that is more accurate, coherent, and of higher quality.

3.4.3. Process-Qualified Accuracy

Although some questions are answered correctly, the reasoning process behind them may not be logically accurate, coherent, or supportive of the final correct answer. As a result, such "hallucination" cases inflate the evaluation of the model's spatial mathematical reasoning ability. To address this, we propose a new composite metric, Process-Qualified

Accuracy (PA), which combines Answer Accuracy and Process Score to more accurately measure the model's true spatial mathematical reasoning capability:

$$\mathcal{PA} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\text{Ans}_{k,i} = \text{GT}_{k,i} \;\; \& \;\; \mathcal{S}_{k,i} \geq 0.75]. \tag{3}$$

We believe that a model can only truly possess the ability to solve a problem when its reasoning process is logically accurate, coherent, and supports the final answer, rather than merely relying on the correctness of the final answer alone.

4. Experiment

DynaSolidGeo supports the random generation of two visual versions: a randomized-view image and a 360-degree rotation video, for each question instance. Since existing geometry problem-solving tasks focus exclusively on the text-image modality, here we also evaluate models under the same text-image setting.

4.1. Experimental Setup

Evaluation Models. We evaluate a range of the latest, popular, and state-of-the-art (SOTA) closed-source and open-source MLLMs. The closed-source models include GPT-5-Nano [27], GPT-5 [27], Gemini-2.5-Flash [11], Gemini-2.5-Pro [11], and Claude-Sonnet-4.5 [5]. The open-source models include LLaVA-OneVision-1.5 family (4B, 8B) [4], GLM-4.5V [17], GLM-4.1V-9B-Thinking [17], Llama-3.2-90B-Vision-Instruct [1], Llama-4-Maverick-17B-Instruct [2], InternVL3-78B [46], InternVL3.5-8B [37], DeepSeek-VL2 [36], and the Qwen3-VL family [30].

Implementation Details. We sample $K = 3$ batches of question instances by setting the *random seed* to 0, 1, and 2, respectively, resulting in a total of 1,509 text-image question instances. For answer evaluation, we allow a 1% relative error tolerance. For process evaluation, we employ Qwen3-14B [39] as the judge model. For the evaluated models, we deploy small-scale models, including the LLaVA-OneVision-1.5 family, GLM-4.1V-9B-Thinking, InternVL3.5-8B, and the Qwen3-VL family (4B, 30B), on NVIDIA A800 GPUs for evaluation. DeepSeek-VL2 is evaluated via the SiliconFlow API[1], while all remaining models are accessed through the OpenRouter API[2] for evaluation.

4.2. Experimental Results

Overall Results on Evaluation Metrics. Table 3 presents the performance of the models in Section 4.1 on the Answer Accuracy (AA), Process Score (PS), and Process-Qualified Accuracy (PA) metrics. For the GPT-5 family, LLaVA-OneVision-1.5 family, and GLM-4.5V, the PS and PA metrics are not reported, as these models either do not disclose their reasoning traces by API or inherently do not produce explicit reasoning processes. Among the closed-source models, GPT-5 achieves the highest overall AA score of 70.8%, outperforming all other models. Among the open-source models, Qwen3-VL-30B-A3B-Thinking attains the highest AA score of 65.6%, surpassing most of the closed-source models. Among the models with available reasoning traces, Qwen3-VL-30B-A3B-Thinking achieves the highest PS and PA scores, both at 65.4%.

Performance Differences across Categories from the Perspective of Spatial Intelligence. As shown in Table 3, the best-performing models perform well in Area calculation (AR), Volume calculation (VC), and Dynamic or moving-point problems (DM). However, all models struggle with Counting problems (CP), and a significant performance gap can be observed between open-source and closed-source models in this category. This discrepancy can be explained through the lens of spatial intelligence theory [14]. Tasks such as AR, VC, and DM mainly rely on lower or mid-level spatial perception, spatial relation, and spatial orientation, where visual cues are explicit and reasoning can be simplified into formula- or rule-based deduction, rather than fine-grained 3D structural reasoning. This aligns well with the representational strengths of current MLLMs, which operate on visual encodings and symbolic reasoning. In contrast, Counting Problems (CP) require higher-order mental rotation and spatial visualization, requiring 3D reconstruction, occlusion reasoning, and mental manipulation of hidden or rotated objects. This explains why models perform relatively well on AR, VC, and DM tasks but fail consistently on CP tasks.

Metric Degradation under Process Evaluation. As shown in Table 3, after introducing process evaluation, all models exhibit varying degrees of decline in both PS and PA metrics compared to AA. Among them, Gemini-2.5-Pro shows the largest drop, with PA decreasing by 9.4% relative to AA, followed by Llama-3.2-90B-Vision-Instruct, whose PA drops by 6%. This suggests that these models, while capable of producing correct answers, often rely on reasoning processes that are less coherent or causally aligned with the final answers. Furthermore, the decline in PA relative to AA is generally smaller for *thinking* models than for *instruct* models. For example, GLM-4.1V-9B-Thinking shows only a 1.5% drop, and Qwen3-VL-8B-Thinking decreases by merely 0.1%, whereas both Llama-3.2-90B-Vision-Instruct and Llama-4-Maverick-17B-Instruct experience drops exceeding 5%. Even within the Qwen3-VL family, Qwen3-VL-8B-Thinking and Qwen3-VL-30B-A3B-Thinking exhibit smaller declines compared to Qwen3-VL-8B-Instruct, Qwen3-VL-30B-A3B-Instruct, and Qwen3-

[1] https://www.siliconflow.com/

[2] https://openrouter.ai/

Model	PD	AN	LC	AR	VC	CP	DM	FP	ALL
	AA / PS / PA	AA / PS / PA	AA / PS / PA	AA / PS / PA	AA / PS / PA	AA / PS / PA	AA / PS / PA	AA / PS / PA	AA / PS / PA
Closed-sourced MLLMs									
GPT-5-Nano	39.5 / - / -	54.0 / - / -	56.1 / - / -	71.9 / - / -	71.2 / - / -	5.7 / - / -	53.0 / - / -	42.6 / - / -	51.4 / - / -
GPT-5	74.6 / - / -	66.0 / - / -	76.8 / - / -	83.6 / - / -	85.3 / - / -	20.0 / - / -	78.8 / - / -	65.1 / - / -	70.8 / - / -
Gemini-2.5-Flash	44.1 / 42.9 / 42.9	48.2 / 45.9 / 45.6	60.1 / 57.8 / 58.1	63.7 / 61.0 / 61.4	55.8 / 53.7 / 53.9	16.2 / 16.2 / 16.2	61.6 / 59.8 / 60.1	34.9 / 33.6 / 33.9	49.6 / 47.8 / 47.9
Gemini-2.5-Pro	71.8 / 61.2 / 54.8	52.4 / 43.6 / 39.2	69.7 / 64.6 / 61.6	71.3 / 67.5 / 66.1	76.9 / 69.7 / 65.4	30.5 / 30.5 / 30.5	63.1 / 60.0 / 57.6	56.4 / 50.4 / 47.7	62.0 / 55.9 / 52.6
Claude-Sonnet-4.5	43.5 / 34.7 / 32.2	26.5 / 25.0 / 24.0	37.9 / 34.5 / 32.8	50.3 / 48.7 / 49.1	53.8 / 49.8 / 47.4	6.7 / 6.0 / 5.7	26.3 / 24.2 / 23.2	15.4 / 13.5 / 12.8	32.7 / 29.7 / 28.6
Open-sourced VLMs									
LLaVA-OneVision-1.5-4B-Instruct	11.9 / - / -	4.5 / - / -	7.6 / - / -	15.8 / - / -	10.3 / - / -	1.9 / - / -	4.0 / - / -	0.0 / - / -	6.8 / - / -
LLaVA-OneVision-1.5-8B-Instruct	17.5 / - / -	1.9 / - / -	7.6 / - / -	2.9 / - / -	2.6 / - / -	1.0 / - / -	5.6 / - / -	3.1 / - / -	5.2 / - / -
GLM-4.5V	49.7 / - / -	31.4 / - / -	42.9 / - / -	57.3 / - / -	50.6 / - / -	7.6 / - / -	48.0 / - / -	12.8 / - / -	38.1 / - / -
GLM-4.1V-9B-Thinking	29.9 / 27.3 / 26.6	22.7 / 21.7 / 21.7	33.3 / 30.7 / 31.3	44.4 / 42.7 / 43.3	41.0 / 37.8 / 39.1	2.9 / 2.9 / 2.9	26.8 / 25.5 / 25.3	5.6 / 4.6 / 4.1	26.2 / 24.6 / 24.7
Llama-3.2-90B-Vision-Instruct	35.6 / 27.1 / 23.2	14.2 / 10.1 / 8.4	25.8 / 22.6 / 21.2	49.7 / 45.6 / 43.9	39.7 / 34.0 / 31.4	1.9 / 1.2 / 1.0	9.1 / 6.7 / 5.6	8.2 / 5.0 / 3.1	22.6 / 18.5 / 16.6
Llama-4-Maverick-17B-Instruct	36.7 / 27.1 / 21.5	13.9 / 10.4 / 8.4	24.7 / 21.8 / 20.7	46.8 / 44.3 / 43.3	38.5 / 31.7 / 30.1	4.8 / 3.3 / 2.9	8.1 / 5.8 / 4.6	8.2 / 5.8 / 4.1	22.1 / 18.2 / 16.3
InternVL3-78B	32.8 / 21.5 / 16.4	3.9 / 3.2 / 2.6	16.7 / 14.3 / 14.1	31.0 / 25.9 / 24.6	22.4 / 18.1 / 18.0	2.9 / 2.4 / 1.9	8.6 / 5.2 / 3.0	4.6 / 2.7 / 1.0	14.6 / 11.0 / 9.6
InternVL3.5-8B	24.3 / 21.6 / 19.8	36.6 / 35.8 / 35.6	33.8 / 33.3 / 33.3	43.9 / 43.0 / 42.1	40.4 / 38.3 / 37.8	7.6 / 6.7 / 6.7	44.4 / 44.1 / 43.9	21.5 / 20.9 / 20.5	33.1 / 32.0 / 31.5
DeepSeek-VL2	10.7 / 5.5 / 1.7	1.0 / 1.0 / 0.3	6.6 / 3.8 / 1.5	12.9 / 9.9 / 7.6	7.1 / 5.3 / 5.1	1.9 / 1.0 / 0.0	2.5 / 1.1 / 0.0	2.6 / 1.0 / 0.0	5.3 / 3.3 / 1.9
Qwen3-VL-8B-Instruct	39.5 / 38.1 / 39.0	48.5 / 47.6 / 48.2	40.4 / 39.8 / 40.4	55.0 / 53.1 / 54.4	50.0 / 49.5 / 50.0	1.9 / 1.7 / 1.9	49.5 / 48.7 / 48.5	30.8 / 29.0 / 29.7	41.9 / 40.8 / 41.4
Qwen3-VL-8B-Thinking	63.3 / 63.3 / 63.3	58.6 / 58.6 / 58.6	59.1 / 58.8 / 59.1	67.8 / 67.7 / 67.8	62.8 / 62.8 / 62.8	9.5 / 9.0 / 8.6	71.2 / 71.1 / 71.2	52.8 / 52.6 / 52.8	58.2 / 58.1 / 58.1
Qwen3-VL-30B-A3B-Instruct	37.3 / 34.2 / 35.0	56.3 / 54.9 / 55.0	54.0 / 52.8 / 54.0	63.2 / 61.7 / 62.0	60.3 / 59.6 / 60.3	6.7 / 6.7 / 6.7	63.6 / 62.6 / 63.1	42.1 / 41.5 / 41.5	50.6 / 49.4 / 49.8
Qwen3-VL-30B-A3B-Thinking	68.4 / 67.8 / 67.8	64.4 / 64.4 / 64.4	67.2 / 67.2 / 67.2	75.4 / 75.1 / 74.9	76.3 / 75.5 / 75.6	11.4 / 11.4 / 11.4	78.3 / 78.2 / 78.3	62.6 / 62.4 / 62.6	65.6 / 65.4 / 65.4
Qwen3-VL-235B-A22B-Instruct	72.3 / 69.9 / 71.8	63.4 / 62.4 / 62.5	65.2 / 64.7 / 65.2	76.0 / 75.1 / 75.4	71.8 / 70.9 / 71.2	6.7 / 6.7 / 6.7	69.7 / 69.2 / 69.2	57.4 / 56.9 / 57.4	63.1 / 62.2 / 62.6

Table 3. Comparison of model performance on the Answer Accuracy (AA), Process Score (PS), and Process-Qualified Accuracy (PA) metrics. For the GPT-5 family, LLaVA-OneVision-1.5 family, and GLM-4.5V, the PS and PA metrics are not reported, as these models either do not disclose their reasoning traces by API or inherently do not produce explicit reasoning processes.

VL-235B-A22B-Instruct. These observations suggest that *thinking* models generally produce reasoning processes that are more coherent, logically sound, and causally consistent with their final answers than those of *instruct* models. In addition, only a few stronger models, such as those from the Gemini and Qwen3-VL family, achieve identical AA, PS, and PA on Counting problems (CP), whereas others show clear metric gaps on different task types. This indicates that the CP task requires higher-order spatial intelligence and more rigorous symbolic reasoning. Consequently, making correct answers less susceptible to hallucination or logical inconsistency.

Data Contamination and Memorization Phenomenon. To probe potential data contamination and memorization effects of VLMs on static datasets, we further evaluate their Answer Accuracy (AA) on the static source questions, as shown in Figure 3. Compared with the static source-question dataset, all models show a notable performance drop on DynaSolidGeo, with Claude-Sonnet-4.5 (-20.4%) and InternVL3.5-8B (-17.6%) declining the most. This reveals that these VLMs may suffer from varying degrees of data contamination on static datasets and tend to rely on memorization-based patterns rather than genuine reasoning processes when producing answers. In contrast, DynaSolidGeo serves as a benchmark for assessing the VLMs' genuine ability in spatial mathematical reasoning. Furthermore, *thinking* models show smaller performance declines than *instruct* ones, suggesting that RL promotes reasoning-oriented behavior, while SFT encourages memorization of answers.

Model	Num. of Output Tokens	Metrics
	Correct / Incorrect / All	AA / PS / PA
Closed-sourced MLLMs		
GPT-5-Nano	8846.12 / 9739.68 / 9280.17	51.4 / - / -
GPT-5	7958.86 / 9900.60 / 8526.33	70.8 / - / -
Gemini-2.5-Flash	4865.65 / 25448.12 / 15231.90	49.6 / 47.8 / 47.9
Gemini-2.5-Pro	15054.69 / 18663.74 / 16425.13	62.0 / 55.9 / 52.6
Claude-Sonnet-4.5	987.64 / 1047.81 / 1028.15	32.7 / 29.7 / 28.6
Open-sourced VLMs		
LLaVA-OneVision-1.5-4B-Instruct	712.62 / 842.79 / 833.90	6.8 / - / -
LLaVA-OneVision-1.5-8B-Instruct	262.11 / 150.33 / 156.19	5.2 / - / -
GLM-4.5V	4679.18 / 5287.26 / 5055.40	38.1 / - / -
GLM-4.1V-9B-Thinking	6599.09 / 7562.91 / 7309.98	26.2 / 24.6 / 24.7
Llama-3.2-90B-Vision-Instruct	756.23 / 821.67 / 806.88	22.6 / 18.5 / 16.6
Llama-4-Maverick-17B-Instruct	761.78 / 866.27 / 843.14	22.1 / 18.2 / 16.3
InternVL3-78B	562.24 / 599.96 / 594.46	14.6 / 11.0 / 9.6
InternVL3.5-8B	11336.19 / 15019.64 / 13801.59	33.1 / 32.0 / 31.5
DeepSeek-VL2	404.41 / 465.09 / 461.88	5.3 / 3.3 / 1.9
Qwen3-VL-8B-Instruct	7514.76 / 18246.12 / 13751.61	41.9 / 40.8 / 41.4
Qwen3-VL-8B-Thinking	13251.57 / 21482.31 / 16693.32	58.2 / 58.1 / 58.1
Qwen3-VL-30B-A3B-Instruct	10236.35 / 22068.35 / 16077.87	50.6 / 49.4 / 49.8
Qwen3-VL-30B-A3B-Thinking	10954.33 / 17192.55 / 13099.88	65.6 / 65.4 / 65.4
Qwen3-VL-235B-A22B-Instruct	6045.59 / 6945.21 / 6370.42	63.1 / 62.2 / 62.6

Table 4. Comparison of average output tokens for correct, incorrect, and overall responses with corresponding performance.

Model Inference Efficiency Analysis. Table 4 summarizes, for each model, the average number of output tokens in the cases of correct and incorrect answers, as well as the overall average, together with the corresponding performance metrics. Overall, the number of output tokens is roughly positively correlated with model performance—models with very poor accuracy consistently pro-

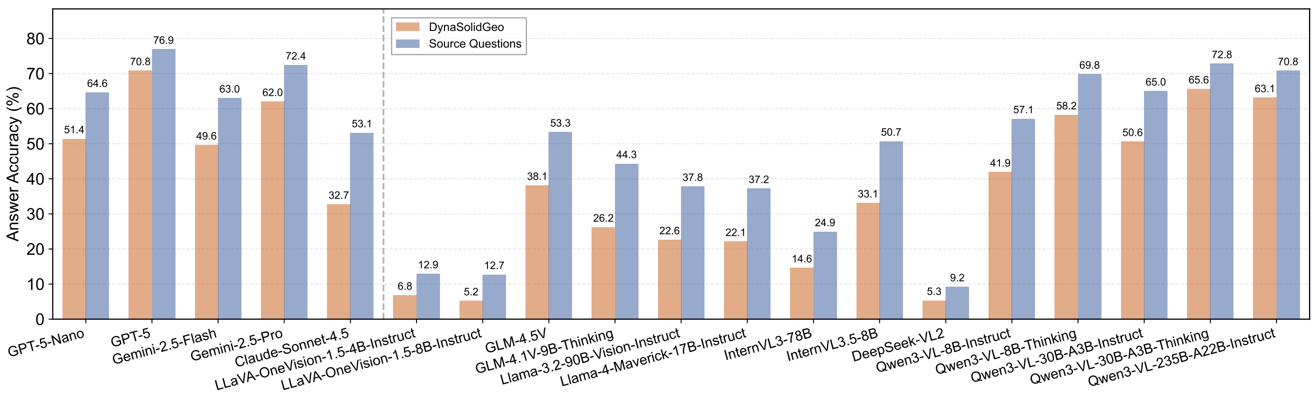

Figure 3. Comparison of model performance on Answer Accuracy (AA) between DynaSolidGeo and source questions.

duce shorter outputs, which aligns with the principle of test-time scaling. In addition, for almost all models (except LLaVA-OneVision-1.5-8B-Instruct), the reasoning traces for incorrect answers are noticeably longer than those for correct ones. This is likely because, when a model encounters a problem it cannot solve or faces logical inconsistencies during reasoning, it tends to repeatedly "rethink" its intermediate steps, resulting in unnecessarily prolonged reasoning chains.

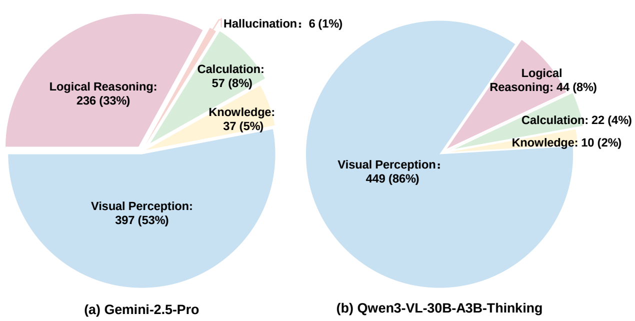

Figure 4. Error Analysis.

4.3. Error Analysis

We conduct an error analysis on two representative models, Gemini-2.5-Pro (closed source) and Qwen3-VL-30B-A3B-Thinking (open source). Specifically, we categorize the errors in Process-Qualified Accuracy (PA) into five types: Visual Perception Errors, Logical Reasoning Errors, Calculation Errors, Knowledge Errors, and Hallucination Errors, as shown in Figure 4. Across the 1,509 sampled instances, Gemini-2.5-Pro makes a total of 715 errors, Qwen3-VL-30B-A3B-Thinking makes 525 errors. Among these error types, visual perception and logical reasoning errors dominate, with visual perception errors accounting for the largest proportion. This indicates that, although these advanced models have demonstrated strong symbolic reasoning capabilities, they still lack sufficient perceptual understanding in solid geometry tasks that require spatial intelligence. Qwen3-VL-30B-A3B-Thinking exhibits 52 more Visual Perception errors than Gemini-2.5-Pro, suggesting that it is more prone to reasoning failures triggered by inaccurate visual perception. This also explains why Qwen3-VL-30B-A3B-Thinking performs significantly worse than Gemini-2.5-Pro on Counting Problems (CP), which demand higher levels of spatial intelligence. In addition, Qwen3-VL-30B-A3B-Thinking makes fewer errors in other categories (including Logical Reasoning Errors), demonstrating its stronger symbolic reasoning capability.

5. Conclusion

In this work, we introduced DynaSolidGeo, the first dynamic benchmark for evaluating the genuine spatial mathematical reasoning capabilities of VLMs in solid geometry. Through a semi-automatic, expert-guided pipeline, DynaSolidGeo enables unbounded generation of diverse multimodal instances, effectively mitigating contamination and memorization issues found in static datasets. By integrating both answer- and process-level evaluation, we provide a more faithful assessment of reasoning validity and spatial perception. Comprehensive experiments uncover persistent limitations in high-level spatial intelligence and reveal substantial performance degradation under dynamic evaluation. We expect DynaSolidGeo to provide a reliable foundation for advancing process-grounded, contamination-resistant multimodal reasoning benchmarks and inspire future research toward robust spatial reasoning in VLMs.

Future Work

We will update our Appendix in the next version.

References

[1] Meta AI. Llama-3.2-vision-instruct. https://huggingface.co/meta-llama/Llama-3.2-90B-Vision-Instruct, 2024. 6

[2] Meta AI. Llama-4-maverick-17b-instruct. https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E-Instruct, 2025. 3, 6

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 3

[4] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025. 1, 3, 6

[5] Anthropic. Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5, 2025. 3, 6

[6] 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. 1

[7] MT Battista. The development of geometric and spatial thinking. *Second handbook of research on mathematics teaching and learning/National Council of Teachers of Mathematics*, 2007. 1

[8] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *arXiv preprint arXiv:2105.14517*, 2021. 1, 2, 3

[9] Simin Chen, Yiming Chen, Zexin Li, Yifan Jiang, Zhongwei Wan, Yixin He, Dezhi Ran, Tianle Gu, Haizhou Li, Tao Xie, et al. Recent advances in large langauge model benchmarks against data contamination: From static to dynamic evaluation. *arXiv preprint arXiv:2502.17521*, 2025. 2

[10] Yuxing Cheng, Yi Chang, and Yuan Wu. A survey on data contamination for large language models. *arXiv preprint arXiv:2502.14425*, 2025. 1

[11] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1, 3, 6

[12] Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models. *arXiv preprint arXiv:2311.09783*, 2023. 1

[13] Yumeng Fu, Jiayin Zhu, Lingling Zhang, Bo Zhao, Shaoxuan Ma, Yushun Zhang, Yanrui Wu, and Wenjun Wu. Geolaux: A benchmark for evaluating mllms' geometry performance on long-step problems requiring auxiliary lines. *arXiv preprint arXiv:2508.06226*, 2025. 2, 3

[14] Howard Gardner. *Frames of mind: The theory of multiple intelligences*. Basic books, 2011. 1, 6

[15] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024. 2, 3

[16] Mary Hegarty and David Waller. Individual differences in spatial abilities. *The Cambridge handbook of visuospatial thinking*, pages 121–169, 2005. 1

[17] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pages arXiv–2507, 2025. 1, 3, 6

[18] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024. 1

[19] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 4999–5007, 2017. 3

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[21] Yucheng Li, Frank Guerin, and Chenghua Lin. An open source data contamination report for large language models. *arXiv preprint arXiv:2310.17589*, 2023. 1

[22] Shijie Lian, Changti Wu, Laurence Tianruo Yang, Hang Yuan, Bin Yu, Lei Zhang, and Kai Chen. Euclid's gift: Enhancing spatial perception and reasoning in vision-language models via geometric surrogate tasks. *arXiv preprint arXiv:2509.24473*, 2025. 1

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 3

[24] 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. 3

[25] 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. 1, 2

[26] Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. *arXiv preprint arXiv:2203.08242*, 2022. 1

[27] OpenAI. Gpt-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf, 2025. Accessed: 2025-10-24. 1, 3, 6

[28] Yonatan Oren, Nicole Meister, Niladri S Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations*, 2023. 1

[29] Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971. 1

[30] Alibaba Cloud / Qwen Team. `https://huggingface.co/collections/Qwen/qwen3-vl`, 2025. 3, 6

[31] Han Wang, Haoyu Li, Brian Ko, and Huan Zhang. On the fragility of benchmark contamination detection in reasoning models. *arXiv preprint arXiv:2510.02386*, 2025. 1

[32] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024. 2, 3

[33] Peijie Wang, Chao Yang, Zhong-Zhi Li, Fei Yin, Dekang Ran, Mi Tian, Zhilong Ji, Jinfeng Bai, and Cheng-Lin Liu. Solidgeo: Measuring multimodal spatial math reasoning in solid geometry. *arXiv preprint arXiv:2505.21177*, 2025. 2, 3, 4

[34] Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, et al. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*, 4, 2024. 2

[35] Xiaobao Wu, Liangming Pan, Yuxi Xie, Ruiwen Zhou, Shuai Zhao, Yubo Ma, Mingzhe Du, Rui Mao, Anh Tuan Luu, and William Yang Wang. Antileakbench: Preventing data contamination by automatically constructing benchmarks with updated real-world knowledge. *arXiv preprint arXiv:2412.13670*, 2024. 2

[36] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 3, 6

[37] Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jianze Liang, Qidong Huang, Jiaqi Wang, Feng Wu, and Dahua Lin. Caprl: Stimulating dense image caption capabilities via reinforcement learning. *arXiv preprint arXiv:2509.22647*, 2025. 3, 6

[38] Liangyu Xu, Yingxiu Zhao, Jingyun Wang, Yingyao Wang, Bu Pi, Chen Wang, Mingliang Zhang, Jihao Gu, Xiang Li, Xiaoyong Zhu, et al. Geosense: Evaluating identification and application of geometric principles in multimodal reasoning. *arXiv preprint arXiv:2504.12597*, 2025. 1, 2, 3, 4

[39] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 6

[40] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 2, 3

[41] Jiaxin Zhang, Zhongzhi Li, Mingliang Zhang, Fei Yin, Chenglin Liu, and Yashar Moshfeghi. Geoeval: benchmark for evaluating llms and multi-modal models on geometry problem-solving. *arXiv preprint arXiv:2402.10104*, 2024. 1, 2, 3

[42] Ming-Liang Zhang, Fei Yin, and Cheng-Lin Liu. A multi-modal neural geometric solver with textual clauses parsed from diagram. *arXiv preprint arXiv:2302.11097*, 2023. 1, 2, 3

[43] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024. 2, 3, 4

[44] Qihao Zhao, Yangyu Huang, Tengchao Lv, Lei Cui, Qinzheng Sun, Shaoguang Mao, Xin Zhang, Ying Xin, Qiufeng Yin, Scarlett Li, et al. Mmlu-cf: A contamination-free multi-task language understanding benchmark. *arXiv preprint arXiv:2412.15194*, 2024. 1

[45] Zihan Zheng, Zerui Cheng, Zeyu Shen, Shang Zhou, Kaiyuan Liu, Hansen He, Dongruixuan Li, Stanley Wei, Hangyi Hao, Jianzhu Yao, et al. Livecodebench pro: How do olympiad medalists judge llms in competitive programming? *arXiv preprint arXiv:2506.11928*, 2025. 1

[46] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 6

[47] Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836*, 2024. 2, 3, 4