



Motivation

⚡ The Gap: From Plane to Solid Geometry

Existing benchmarks are stuck in plane geometry. We target the overlooked challenge of solid geometry to measure true spatial reasoning.

⚡ The Target: Complex Spatial Math Reasoning

Solid geometry provides a testbed for spatial intelligence—the ability to perform complex mathematical reasoning in a 3D world.

⚡ Forging the Path to AGI:

Challenging models with complex 3D tasks fosters the advanced reasoning essential for the next leap toward AGI.

Statistics of SOLIDGEO

Statistic	Number
Total questions	3,113
- Multiple-choice questions	969
- Open-ended questions	2144
- Single-step questions	1936
- Multi-step questions	208
Difficulties (Easy: Medium: Hard)	32%:61%:7%
Newly collected questions	1,376 (44.2%)
Existing-dataset questions	1,737 (55.8%)
Newly collected images	3,555 (66.1%)
Existing-dataset images	1,825 (33.9%)
Language	
- English (EN)	2,192 (70.4%)
- Chinese (CN)	921 (29.6%)
Maximum question length	679
Maximum answer length	2833
Average question length	77.2
Average answer length	312.2

Table 1: Key Statistics of SOLIDGEO.

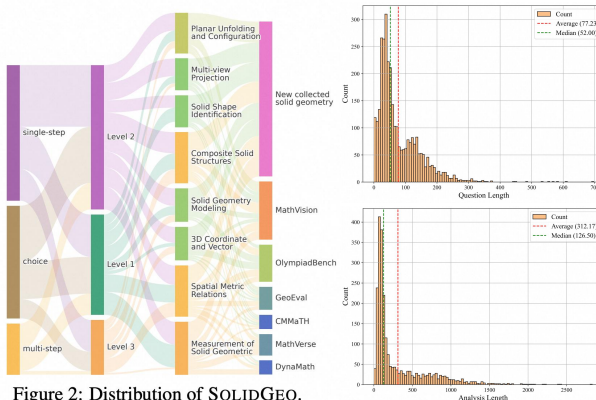


Figure 2: Distribution of SOLIDGEO.

Table 2: Comparison with existing multimodal math benchmarks. SG: Solid Geometry, PG: Plane Geometry. **Level:** **K**=K-12, **U**=University, **C**=Competition. **Source:** **S**=Self-sourced, **P**=Collected from Public Dataset. MC: Multiple Choice, SS: Single-Step, MS: Multi-Step.

Benchmarks	Language	SG Size	SG Proportion	PG Proportion	Level	Source	SG category	MC	SS	MS
GeoQA [31]	CN	0	0.0%	100.0%	K	S	X	✓	X	X
Geometry3K [30]	EN	0	0.0%	100.0%	K	S	X	✓	✓	X
UniGeo [33]	EN	0	0.0%	100.0%	K	S	X	✓	✓	X
PGPS9K [32]	EN	0	0.0%	100.0%	K	S	X	✓	✓	X
MMMU-MATH [38]	EN	0	0.0%	29.3%	U	S	X	✓	✓	X
GeoEval [5]	EN	100	2.0%	94.1%	K	P	X	✓	X	X
DynaMath [61]	EN	150	3.0%	15.4%	K	S	X	✓	✓	X
MATH-Vision [6]	EN	263	8.7%	58.7%	KU	S	X	✓	✓	X
OlympiadBench [62]	EN/CN	784	9.2%	15.6%	C	S	X	✓	✓	X
MathVerse [7]	EN	119	15.1%	64.7%	K	SP	X	✓	X	X
GeoSense [41]	EN/CN	350	20.0%	80.0%	K	SP	2	✓	✓	X
SOLIDGEO (Ours)	EN/CN	3113	100.0%	0.0%	KC	SP	8	✓	✓	✓

Overview and Main Result

Multipile Choice	Open-ended: Single-Step	Open-ended: Multi-Step
<p>Question: Which of the cones listed below can be formed from a 252° sector of a circle of radius 10 by aligning the two straight sides?</p> <p><image1> <image2></p> <p>A. A B. B C. C D. D E. E</p>	<p>Question: A regular icosahedron is a 20-faced solid where each face is an equilateral triangle and five triangles meet at every vertex. The regular icosahedron shown below has one vertex at the top, one vertex at the bottom, an upper pentagon of five vertices all adjacent to the top vertex and all in the same horizontal plane, and a lower pentagon of five vertices all adjacent to the bottom vertex and all in another horizontal plane. Find the number of paths from the top vertex to the bottom vertex such that each part of a path goes downward or horizontally along an edge of the icosahedron.</p>	<p>Question: As shown in Figure 1, this is a Rubik's Cube composed of 8 equally sized smaller cubes, with a total volume of 64.</p> <p>(1) Determine the edge length of the Rubik's Cube.</p> <p>(2) In the figure, quadrilateral ABCD is a square. Find the area and the side length of this square.</p> <p>(3) Place the square ABCD on the number line as shown in Figure 2, such that point A coincides with -1, point E coincides with 1, and point F is symmetric to point D with respect to point E. Then, the number represented by point D on the number line is .</p> <p><ImageHere> Figure 1 <ImageHere> Figure 2</p>
<p>Answer: C</p> <p>Difficulty Level: 2</p> <p>Problem Type: PUC</p>	<p>Answer: 810</p> <p>Difficulty Level: 3</p> <p>Problem Type: CSS</p>	<p>Answer: (1) 4 (2) 8 2√2 (3) -1-2√2</p> <p>Difficulty Level: 2</p> <p>Problem Type: CSS</p>

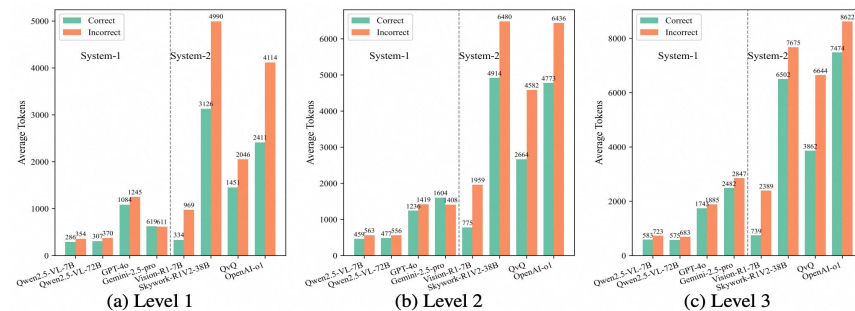
Model	Overall	CSS	SMR	SSI	PUC	MSGF	SGM	MVP	3DCV	Avg.tokens
Text-only, zero-shot direct answering										
Deepseek-V3[12] (LLM)	9.3	10.7	8.1	8.3	12.7	6.3	7.8	10.3	12.2	787.2
GPT-4o[26] (MLLM)	9.1	10.0	10.4	10.6	6.8	12.1	8.6	7.3	9.6	692.6

Open-source MLLMs (Text + Image, zero-shot direct answering)										
System-1 Models										
LLaVA-v1.5-7B[13]	1.8	1.1	1.1	6.7	2.2	0.6	0.0	4.6	0.0	246.2
InternLM-XComposer2.5-VL-7B[63]	4.4	2.5	1.8	6.7	8.9	0.6	0.0	9.4	1.2	151.8
DeepSeek-VL2-7B[64]	5.1	2.8	2.6	11.1	5.1	1.4	1.8	11.7	1.8	338.2
Math-LLaVA-13B[65]	5.9	4.2	4.1	7.6	11.7	2.7	4.2	12.6	6.2	7.4
LLaVA-NeXT-Interleave-7B[66]	7.7	2.5	2.3	21.5	13.5	2.3	7.3	16.7	0.6	486.3
LLaVA-OneVision-Chat-7B[47]	8.6	4.3	2.9	19.3	15.2	3.5	6.4	17.9	0.0	353.2
Qwen2.5-VL-Instruct-7B[28]	15.5	8.4	8.8	30.1	13.3	26.2	16.2	15.2	10.2	490.2
LLaVA-OneVision-Chat-72B[47]	15.9	13.2	9.5	31.9	18.1	12.9	11.8	23.7	8.4	396.3
InternVL3-8B[15]	17.7	11.8	10.0	24.4	17.4	28.0	19.1	19.9	7.2	488.8
Mistral-small-3.1-24b-instruct[67]	19.6	15.2	15.8	27.4	17.1	28.9	10.9	17.0	16.8	769.7
Qwen2.5-VL-Instruct-72B[28]	24.2	19.7	18.8	29.6	21.5	35.4	16.4	22.5	18.0	485.0
InternVL3-78B[15]	26.2	17.4	17.9	34.8	24.9	36.8	22.7	30.5	17.4	493.2
Llama-4-Maverick-17B-128E[36]	29.6	25.1	30.9	34.6	20.5	43.4	32.6	20.7	26.3	605.6
System-2 Models										
LlamaV-o1-11B [68]	1.5	0.6	0.7	1.5	0.5	5.0	2.7	0.1	0.0	106.1
LLaVA-CoT-11B [69]	7.3	4.2	2.5	7.4	6.5	15.1	8.2	7.4	1.8	401.7
VLM-R1-3B [70]	9.6	6.3	4.4	11.1	8.7	19.6	4.5	8.3	2.4	453.0
R1-Onevision-7B [71]	13.2	7.7	9.7	25.2	10.1	23.3	11.8	12.3	9.0	522.3
Vision-R1-7B [72]	18.1	11.7	11.3	28.6	17.8	26.9	13.9	19.3	12.0	1498.7
Skywork-R1V2-38B [73]	23.0	18.4	29.5	13.3	11.6	31.2	30.0	12.3	26.9	5682.9
QvQ-72B-Preview [74]	26.6	17.9	28.1	37.0	22.9	34.7	20.9	20.3	22.8	3622.2

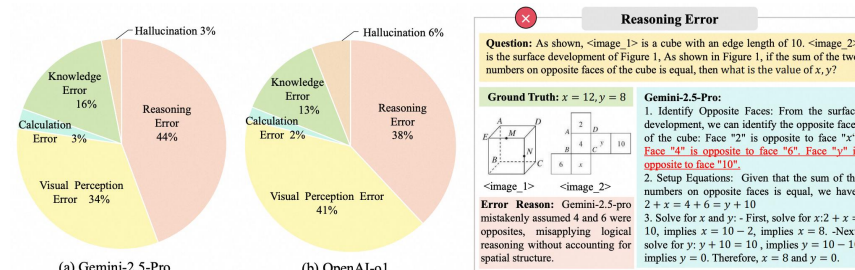
Closed-source MLLMs (Text + Image, zero-shot direct answering)										
System-1 Models										
Claude-3.5-Sonnet[46]	22.2	16.9	9.8	42.2	24.2	36.5	25.5	23.5	9.6	992.1
GPT-4V[14]	25.3	16.6	15.8	35.6	21.5	41.5	25.5	25.9	18.0	1433.5
Gemini-1.5-pro[27]	25.3	18.5	16.8	34.8	19.6	41.6	17.3	25.6	19.2	1003.5
GPT-4o[26]	25.5	18.9	16.8	32.6	19.6	41.0	17.3	26.5	19.2	1344.9
Claude-3.7-Sonnet[46]	34.1	27.7	28.2	43.0	32.9	46.8	43.6	28.5	26.3	1217.4
Gemini-2.5-pro[27]	42.7	52.0	75.7	24.8	20.9	26.0	58.4	19.6	72.9	1263.9
System-2 Models										
OpenAI-o1[75]	49.5	48.7	54.2	48.9	36.1	55.3	59.1	43.0	55.1	4942.6
Human Performance										
Human	77.5	88.2	70.9	90.2	77.2	87.4	71.2	78.5	69.2	-

Fine-grained Analysis

Models	Original	CoT	CoT & 2-shot	Level 1	Level 2	Level 3	MC	SS	MS
Claude-3.5-sonnet [46]	22.2	21.7(+0.5)	19.3(+1.9)	37.7	15.7	4.5	32.8	17.8	13.9
Gemini-1.5-pro [27]	25.3	26.1(+0.8)	27.5(+2.2)	39.0	19.7	8.5	32.1	23.1	14.4
GPT-4V [14]	25.3	24.1(+1.2)	23.4(+1.9)	41.6	18.5	6.8	35.2	22.1	10.6
GPT-4o [26]	25.5	24.9(+0.6)	22.9(+2.6)	38.9	20.1	7.9	32.7	23.2	13.0
Claude-3.7-Sonnet [46]	34.1	-	-	42.5	31.2	16.4	39.9	32.5	21.2
Gemini-2.5-pro [27]	42.7	-	-	22.1	50.8	80.7	35.2	49.0	19.3
OpenAI-o1 [75]	49.6	-	-	46.6	50.4	57.4	50.8	51.2	28.8
LLaVA-OneVision-Chat-7B [47]	8.6	10.5(+1.9)	11.4(+2.8)	16.1	5.4	1.1	22.1	2.6	1.4
Qwen2.5-VL-Instruct-7B [28]	15.5	16.0(+0.5)	15.0(+0.5)	28.1	10.0	2.8	22.2	13.1	6.2
LLaVA-OneVision-Chat-72B [47]	15.9	14.8(+1.1)	15.6(+0.3)	23.5	13.0	4.0	29.2	10.2	6.2
InternVL3-8B [15]	17.7	18.1(+0.4)	18.3(+0.6)	30.7	11.9	6.2	27.3	14.1	6.2
Qwen2.5-VL-Instruct-72B [28]	24.2	28.8(+4.6)	26.1(+1.9)	33.5	20.6	10.6	28.5	22.8	17.3
InternVL3-78B [15]	26.2	25.8(+0.4)	25.9(+0.3)	42.2	19.7	6.2	35.1	23.0	15.4
QvQ-72B-Preview [74]	26.6	-	-	32.6	24.5	14.7	31.5	26.3	6.7



Error Analysis and Case Study



Summary

- 🚀 Identify a Critical Gap: Current benchmarks neglect solid geometry, a crucial domain for spatial reasoning required for AGI.
- 🚀 Develop SOLIDGEO: The first benchmark dedicated to solid geometry, featuring 3113 problems with detailed annotations.
- 🚀 Reveal a Major Performance Gap: Evaluated 27 leading MLLMs, uncovering their significant limitations in 3D reasoning and a substantial gap compared to human performance.