

Do MLLMs Really Understand Space? A Mathematical Spatial Reasoning Evaluation

Shuo Lu^{1,2*,†}, Jianjie Cheng^{3*,†}, Yinuo Xu^{1,†}, Yongcan Yu^{1,†},
Lijun Sheng¹, Peijie Wang¹, Siru Jiang¹, Yongguan Hu¹, Lingxiao He³,
Meng Wang³, Qianlong Xie³, Xingxing Wang³, Ran He^{1,2}, Jian Liang^{1,2‡}

¹ NLPR & MAIS, Institute of Automation, Chinese Academy of Sciences (CASIA)

² School of Artificial Intelligence, University of Chinese Academy of Sciences ³ Meituan Inc.

 <https://github.com/shuolucs/MathSpatial> {shuolucs, liangjian92}@gmail.com

Abstract

Multimodal large language models (MLLMs) have achieved strong performance on perception-oriented tasks, yet their ability to perform mathematical spatial reasoning, defined as the capacity to parse and manipulate two- and three-dimensional relations, remains unclear. While humans easily master textbook-style spatial problems, current MLLMs exhibit substantially higher error rates, exposing a fundamental weakness in spatial comprehension. To investigate this gap, we present *MathSpatial*, a controlled diagnostic benchmark and training resource for mathematical spatial reasoning in MLLMs. *MathSpatial* provides two complementary subsets: *MathSpatial-Bench* (2,000 problems) uses minimal-noise educational diagrams to isolate spatial reasoning from perceptual confounds, and *MathSpatial-Corpus* (8,000 problems) provides verified solutions and detailed operation-typed structured reasoning traces for training. Benchmarking 20+ MLLMs reveals a severe deficit: non-reasoning models largely languish below 60% accuracy, and even the strongest reasoning models fall ~16 percentage points behind human performance (96.3%). Crucially, we demonstrate that fine-tuning on *MathSpatial-Corpus* yields consistent performance gains across model families. At the 10,000-example scale, *MathSpatial* provides a practical resource for diagnosing and advancing spatial reasoning with clean, costly-to-curate supervision.

1 Introduction

Multimodal large language models (MLLMs) have advanced rapidly on perception-oriented tasks (Sarto et al., 2025; Kuang et al., 2025; Ye et al., 2025). However, whether MLLMs can truly perform spatial reasoning, the ability to parse and

manipulate 2D and 3D relationships, remains an open question. This ability is central to higher-level cognition and underpins applications such as robotic manipulation (Yuan et al., 2025), autonomous driving (Tian et al., 2025), and embodied intelligence (Feng et al., 2025), yet we find current models frequently fail on problems that humans solve with ease, as shown in Figure 1(a).

Existing research (Zha et al., 2025; Tang et al., 2025) on spatial reasoning has made notable progress, yet three challenges remain (Figure 1b). First, existing benchmarks intentionally span natural scenes, rendered environments, and educational diagrams (Ma et al., 2025; Wang et al., 2024; Jia et al., 2025; Zhang et al., 2024). This breadth is valuable, but differences in visual complexity, supervision, task format, and scoring can make reasoning errors difficult to separate from perceptual failures. Second, few resources combine diagnostic evaluation with large-scale, verified training supervision for mathematical spatial reasoning, as summarized in Table 1. Third, most methods treat spatial reasoning as an end-to-end black-box mapping (Chen et al., 2024a; Ray et al., 2024), without interpretable intermediate supervision, making it difficult to diagnose failure modes or provide targeted improvements.

Building on these insights, we introduce *MathSpatial*, a controlled diagnostic benchmark and training resource for mathematical spatial reasoning in MLLMs. Grounded in simple yet diagnostic educational problems, *MathSpatial* advances the field along three dimensions: *controlled evaluation* through minimal-noise diagrams, *large-scale training data* with verified solutions, and *structured annotations* that support interpretable reasoning.

To mitigate *perceptual confounds*, *MathSpatial-Bench* provides 2,000 evaluation problems using clean geometric diagrams with minimal background and texture, isolating reasoning from visual noise. It spans three categories—Holistic

*Equal contribution.

†Core contributors.

‡Corresponding author.

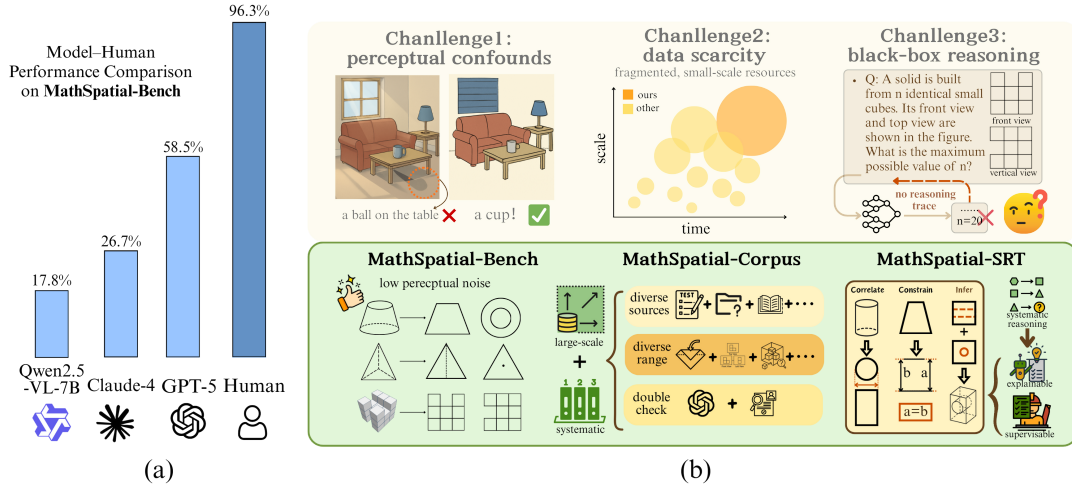


Figure 1: Left: On *MathSpatial-Bench*, humans achieve over 95% accuracy while most non-reasoning MLLMs remain below 60%; even GPT-5 falls short of human-level performance. Right: Three core-challenges of spatial reasoning and the design of *MathSpatial* to address them.

Recognition, Generative Inference, and Abstract Deduction—across 11 subtypes, enabling fine-grained diagnostic analysis. On this benchmark, humans achieve 95%+ accuracy while most non-reasoning MLLMs remain below 60%, and even the strongest reasoning model reaches only $\sim 80\%$. To address *data scarcity*, *MathSpatial-Corpus* offers 8,000 training problems sourced from textbooks and exam banks spanning primary to high school levels, covering tasks from three-view recognition to geometric property calculation, each with human-verified solutions. To enable *interpretable reasoning*, each problem is further annotated with structured reasoning traces composed of three atomic operations (CORRELATE, CONSTRAIN, and INFER) providing explicit intermediate supervision beyond black-box training. All data undergoes multi-stage quality control including geometric consistency checking, cross-split deduplication, and solution verification.

In summary, our contributions are threefold:

- We collect and release *MathSpatial*, a 10,000-example resource of high-quality spatial reasoning problems with multi-stage quality control. It consists of *MathSpatial-Bench* (2K problems across 3 categories and 11 subtypes) for controlled diagnostic evaluation and *MathSpatial-Corpus* (8K problems with human-verified solutions) for training.
- We provide structured reasoning annotations for all instances within the *MathSpatial-Corpus*, grounded in three atomic operations (CORRELATE, CONSTRAIN, and INFER). This

yields interpretable intermediate supervision to catalyze future research in spatial reasoning.

- We benchmark 20+ leading MLLMs on *MathSpatial-Bench*. Conventional models struggle to surpass 60% accuracy, and even the most advanced reasoning models lag behind human performance (96.3%). Furthermore, our in-depth error analysis identifies geometric rule violations and logical reasoning gaps as the predominant failure modes hindering further progress.

2 Related Work

2.1 MLLM Spatial Reasoning

Multimodal large language models (MLLMs) integrate textual and visual modalities and have demonstrated remarkable potential across vision-language tasks (Li et al., 2023; Chen et al., 2024b; Bai et al., 2025; Phan et al., 2025). Recent studies indicate that MLLMs exhibit emerging spatial reasoning abilities, such as object-relation understanding, mental rotation, and geometric transformation (Yang et al., 2025; Li et al., 2025a; Wang et al., 2025b), typically enhanced through supervised fine-tuning on curated datasets, whose role in reasoning-oriented vision-language models has been examined in both survey and empirical studies; parallel multimodal-generation work studies human-feedback and preference alignment under diverse user groups (Yu et al., 2025; Lu et al., 2025a, 2026a). Recent work also explores geometric imagination for spatial reasoning from limited views (Chen et al., 2025), while target-aware

Table 1: Comparison with existing spatial/geometry reasoning benchmarks. *Vis. Noise*: Minimal/Moderate/Heavy. *Source*: R/S/E denote real/synthetic/educational data. *Supervision*: “-” denotes evaluation-only, and SRT denotes operation-typed traces.

Dataset	Source	Vis. Noise	#Tasks	#Samples	Answer	Supervision	Scoring
SpatialVLM (Chen et al., 2024a)	R-WebLi	Heavy	2	546	Mix	Auto	Numeric tol.
GeoEval (Zhang et al., 2024)	E-Edu	Moderate	3	5,050	Mix	-	Exact match
3DSRBench (Ma et al., 2025)	R+S-COCO	Heavy	4	6,942	Mix	-	Exact match
SOLIDGEO (Wang et al., 2025a)	E-Edu	Moderate	8	3,113	Mix	-	Exact match
VSI-Bench (Yang et al., 2025)	R-Indoor	Heavy	8	288	Mix	-	Numeric tol.
OmniSpatial (Jia et al., 2025)	R-Internet	Heavy	50	1,387	Closed	-	LLM judge
MathSpatial (Ours)	E-Edu	Minimal	11	10,000	Closed	Ans+Sol+SRT	Exact+Human

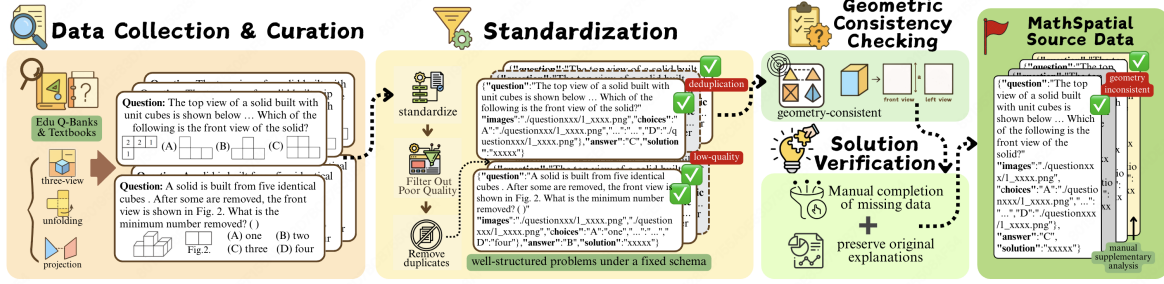


Figure 2: *MathSpatial* source data construction pipeline: Data Collection and Curation → Standardization → Geometric Consistency Checking → Solution Verification.

calibration data selection studies highlight the importance of preserving uncertainty during model adaptation (Yang et al., 2026b). For instance, SpatialVLM (Chen et al., 2024a) and SAT (Ray et al., 2024) leverage geometric labels and spatial rules, while MM-Spatial (Daxberger et al., 2025) and Spatial-MLLM (Wu et al., 2025) employ multi-view 3D scenes and large-scale corpora to strengthen cross-view reasoning. Complementary multimodal research studies physical dynamics and interaction in 3D scenes, while knowledge distillation provides a route toward efficient transfer to lightweight models (Li et al., 2026a,b, 2025b). However, these approaches rely on free-form CoT supervision and lack large-scale, systematically curated training data for mathematical spatial reasoning, a gap that *MathSpatial-Corpus* directly addresses. Inference-time reasoning elicitation via reinforcement-based query refinement offers another complementary direction (Zhou et al., 2026). Adjacent vision work has likewise studied uncertainty-guided active sampling, self-correcting visual agents, and consistency-based zero-shot inference (Zhang et al., 2026; Yang et al., 2026a; Wang et al., 2026); these methods target restoration or segmentation, whereas our focus is mathematical spatial reasoning. For distributed deployment, communication-efficient asynchronous federated

learning and privacy-preserving federated analytics have been explored for visual recognition (Hou et al., 2026b,a); *MathSpatial* instead studies centralized diagnostic evaluation and structured reasoning supervision.

2.2 Spatial Reasoning Benchmarks

Spatial reasoning benchmarks differ substantially in their goals and construction. SpatialVLM (Chen et al., 2024a) and 3DSRBench (Ma et al., 2025) emphasize real or synthetic scenes with substantial visual complexity; adjacent physically grounded 3D-world synthesis benchmarks broaden the scope beyond question answering (Lu et al., 2026b). VSI-Bench (Yang et al., 2025) and OmniSpatial (Jia et al., 2025) broaden the task range in indoor or Internet imagery. GeoEval (Zhang et al., 2024) and SOLIDGEO (Wang et al., 2025a) instead use educational geometry and therefore provide a closer comparison to our setting. These resources also vary in answer format, supervision, and evaluation protocol, from exact matching or numeric tolerance to LLM-based judging. Table 1 summarizes these dimensions rather than treating all spatial benchmarks as directly interchangeable. These differences also relate to broader concerns about trustworthiness and reliable evaluation in reasoning systems (Wang et al., 2025c; Lu et al., 2025b). Re-

cent mathematical reasoning studies further explore adversarial constraint graph synthesis, diagnostic iterative alignment, and exploration-efficient iterative alignment (Yu et al., 2026; Chen et al., 2026; Rao et al., 2026), reinforcing the importance of data quality and diagnostic supervision. *MathSpatial* occupies a complementary position: its minimal-noise educational diagrams are deliberately controlled to diagnose mathematical spatial reasoning, while its 8K training split supplies verified answers, solutions, and operation-typed SRT supervision. The contribution is thus the combination of controlled diagnosis and scalable structured supervision, not greater visual complexity or broader open-ended task coverage.

3 MathSpatial

3.1 Overview

In this section, we present *MathSpatial*, a large-scale dataset for evaluating and advancing mathematical spatial reasoning. It comprises 10K problems: 2K from *MathSpatial-Bench* for diagnostic evaluation and 8K from *MathSpatial-Corpus* for systematic training. Beyond standard problem-solution pairs, every item in the corpus is annotated with structured reasoning traces (*MathSpatial-SRT*) organized around three atomic operations (CORRELATE, CONSTRAIN, INFER), providing intermediate supervision that goes beyond conventional answer-only or free-form CoT annotations.

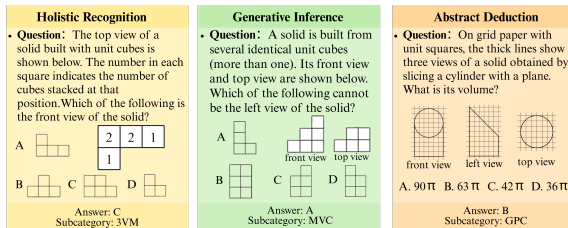


Figure 3: Examples of three *MathSpatial* categories.

3.2 MathSpatial Data Collection

To construct *MathSpatial*, we design a semi-automated pipeline (Figure 2) that ensures both large scale and high quality.¹

- **Collection.** We collect spatial reasoning problems from diverse educational sources across grades, regions, and textbook editions, focusing on objective answers. From 35,428 raw

¹All problems are derived from public educational materials. Copyright-sensitive items are excluded.

candidates, an initial curation step filters structurally incomplete and non-spatial items, retaining 21,673 problems.

- **Standardization.** All problems are converted into a unified schema (image, question, choices, answer, solution). Cross-split de-duplication (MD5 hashing + GPT-4.1 visual similarity + semantic filtering) and manual verification remove duplicates and low-quality items, yielding ~ 11 K unique samples. Chinese problems are translated into English with spot-checked quality.
- **Geometric Consistency Checking.** Rule-based verification with human-in-the-loop review checks cross-view correspondence, dashed/solid line conventions, and orthographic projection rules. Approximately 0.4K inconsistent items are removed; remaining cases without annotator consensus are discarded from the dataset.
- **Solution Verification.** Problems with official solutions are retained directly. For the ~ 0.8 K items lacking solutions, graduate students trained in geometry derive solutions with rigorous dual-reviewer cross-validation to ensure correctness.
- **Final Output.** After all stages, 10,000 verified problems remain, partitioned into *MathSpatial-Bench* (2K) for evaluation and *MathSpatial-Corpus* (8K) for training. As shown in Table 1, *MathSpatial* combines a controlled spatial-reasoning focus with large-scale answer, solution, and SRT supervision.

3.3 MathSpatial-Bench

Building on the partition introduced above, we focus on *MathSpatial-Bench*, the 2K-problem benchmark specifically designed to evaluate spatial reasoning in multimodal large language models. Rather than emphasizing scale, the benchmark prioritizes clarity and diagnostic value: problems are selected to minimize perceptual distractions and to isolate reasoning skills such as cross-view correspondence, geometric consistency, and deductive inference under strict conditions.

MathSpatial-Bench covers three major categories, further divided into eleven subcategories that reflect representative tasks in human education. The three categories are Holistic Recognition, Generative Inference, and Abstract Deduction. Figure 4 illustrates representative examples from each category, showing how tasks require different forms of reasoning such as multi-view correspondence, geometric consistency, and property deduction.

In total, the benchmark consists of 518 prob-

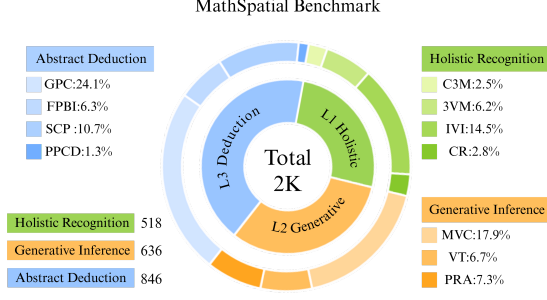


Figure 4: *MathSpatial-Bench* distribution.

lems in Holistic Recognition, 636 in Generative Inference, and 846 in Abstract Deduction. The detailed distribution across all subcategories is shown in Figure 4, where for instance Geometric Property Calculation (GPC) accounts for 24.1% and Missing-View Completion (MVC) for 17.9%. The balance enables fine-grained error analysis and a comprehensive assessment of model capabilities.

3.4 MathSpatial-Corpus

MathSpatial-Corpus provides 8,000 problems for training, each comprising a problem image, textual description, final answer, and detailed solution. A portion of the problems originally in Chinese is translated into English, enabling bilingual support. Together with *MathSpatial-Bench*, it completes the evaluation-training loop of *MathSpatial* and provides substantial task-aligned supervision for spatial reasoning.

Structured Reasoning Annotations. Beyond standard problem-solution pairs, every item is annotated with structured reasoning traces organized around three atomic operations: CORRELATE (establishing cross-view correspondences between geometric entities), CONSTRAIN (applying geometric and projection rules to enforce consistency), and INFER (deducing latent attributes or final answers). Concretely, given a problem (x, y) , we use GPT-4o (OpenAI, 2024) to generate operation-level traces $r = (r_1, \dots, r_T)$ under a constrained schema specifying operation type, arguments, and assertion. To ensure annotation quality, each trace is validated through a dual-role review scheme: a Reviewer agent audits every step for operation-type errors, contradictions, or missing steps, and a Checker agent rewrites the trace to correct identified issues while preserving the schema. This process detects and fixes $\sim 10\%$ of generated traces.

Generation and training. Given a problem (x, y) , we use GPT-4o to generate operation-level traces

$r = (r_1, \dots, r_T)$ under a constrained schema (operation type + arguments + assertion). We linearize (r, y) into a single sequence and apply supervised fine-tuning. This strategy enables models to learn not only to predict correct answers but also to produce interpretable, verifiable reasoning traces that reflect the atomic operation structure. To reduce structural or logical noise, each SRT is validated through a dual-role scheme. In this process, a Reviewer agent audits every CORR/CONS/INFER step for operation-type errors, contradictions, or missing steps, and a Checker agent rewrites the trace to correct all identified issues while preserving the SRT schema.

Corpus Statistics. The corpus spans 11 subtypes with a K-12 skew: GPC dominates (65%), while rare tasks (e.g., CR) are $<1\%$. It contains $\sim 20K$ images (2.6/problem) in mixed formats (46% choice, 54% fill-in). Crucially, all bilingual problems feature structured reasoning traces (avg. 5.3 steps) for robust supervision. **Qualitative example.** To illustrate the difference between free-form reasoning and structured traces, Figure 5 compares model outputs on the same problem. GPT-4o (OpenAI, 2024) produces a partially correct but logically inconsistent explanation, while Qwen2.5-VL-7B (Bai et al., 2025) yields fragmented reasoning with calculation mistakes. In contrast, *MathSpatial-SRT* generates a clear sequence of CORRELATE, CONSTRAIN, and INFER steps, leading to the correct answer with transparent intermediate evidence. This highlights how structured supervision not only improves accuracy but also yields interpretable reasoning traces.

4 Experiments

4.1 Experimental Setup

We evaluate a broad range of MLLMs on *MathSpatial-Bench* to assess the benchmark’s diagnostic capability across model families. Closed-source systems include GPT-5.4/5/4.1/4o (OpenAI, 2026, 2025, 2024), Gemini-3.1-pro/3-flash/2.5-pro/2.5-flash (Gemini Team, 2026, 2025b,a), and Claude Opus 4.6/Sonnet 4.6/Sonnet 4/3.7/3.5 (Anthropic, 2026, 2025b,a, 2024). Open-source baselines include Qwen2.5-VL-7B/72B (Bai et al., 2025), InternVL3-8B (Zhu et al., 2025), Llama-4-Maverick-17B (Singh, 2025), Kimi-VL-A3B (Team et al., 2025), GLM-4.5V (Hong et al., 2025), DeepSeek-VL2-7B (Wu et al., 2024), Pixtral-12B (Agrawal et al., 2024), Idefics3-8B (Laurençon et al., 2024), and Llama3-

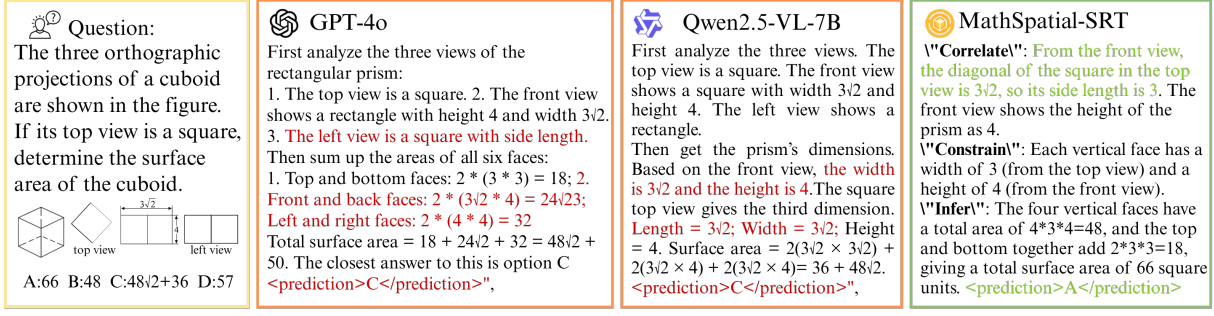


Figure 5: Qualitative comparison of model outputs on a sample problem. GPT-4o and Qwen2.5-VL-7B produce free-form reasoning with inconsistencies or errors, *MathSpatial-SRT* yields a structured reasoning trace based on atomic operations (CORRELATE, CONSTRAIN, INFER), leading to a correct solution.

Table 2: Evaluation on *MathSpatial-Bench*. Accuracy (%) per subtype and average token usage. Red and blue indicate the best and second-best LMMs per column. Gray indicates models fine-tuned on *MathSpatial-Corpus*. † indicates reasoning (thinking) models. Overall results for four key models report mean±standard deviation over three runs.

Model	Holistic Recognition				Generative Inference			Abstract Deduction				Overall	Avg. Token
	C3M	3VM	IVI	CR	MVC	VT	PRA	GPC	FPBI	SCP	PPCD		
Closed-Source Models													
GPT-5.4 (OpenAI, 2026)	51.0	62.6	61.7	55.4	58.3	41.8	69.7	40.5	50.4	59.2	65.4	54.2±1.1	333.2
GPT-5 (OpenAI, 2025)	28.6	53.7	66.2	48.2	61.6	44.8	71.7	52.3	60.0	68.1	57.7	58.5	676.3
GPT-4.1 (OpenAI, 2025)	2.0	45.5	43.1	26.8	24.1	0.0	55.2	0.0	29.6	18.3	46.2	22.6	676.3
GPT-4o (OpenAI, 2024)	2.0	36.6	34.1	33.9	23.5	1.5	45.5	0.0	26.4	16.4	26.9	19.6	677.4
Gemini-3.1-pro (Gemini Team, 2026) [†]	71.4	74.0	78.6	66.1	80.4	72.4	88.3	70.7	68.8	73.8	80.8	75.2±0.6	7516.9
Gemini-3-flash (Gemini Team, 2025b) [†]	73.5	81.3	83.4	71.4	84.0	79.1	85.5	75.9	76.8	76.1	88.5	79.8	10720.7
Gemini-2.5-pro (Gemini Team, 2025a)	31.4	42.5	43.5	26.7	58.5	44.9	28.1	45.6	41.7	49.6	33.3	44.9	913.8
Gemini-2.5-flash (Gemini Team, 2025a)	34.3	42.5	46.7	26.7	58.5	47.8	43.8	50.0	25.0	57.4	33.3	48.5	1115.2
Claude Opus 4.6 (Anthropic, 2026)	44.9	57.7	59.0	50.0	44.5	38.8	69.7	36.7	50.4	49.3	57.7	47.8±0.9	964.2
Claude Sonnet 4.6 (Anthropic, 2026)	42.9	55.3	59.7	39.3	46.2	31.3	74.5	37.1	52.8	44.6	69.2	47.9	941.7
Claude Sonnet 4 (Anthropic, 2025b)	6.1	33.3	46.2	21.4	30.8	1.5	60.7	0.2	32.0	41.3	53.8	26.7	1005.5
Claude Sonnet 3.7 (Anthropic, 2025a)	2.0	36.6	39.3	21.4	21.6	1.5	57.9	0.0	23.2	26.3	34.6	21.5	885.8
Claude Sonnet 3.5 (Anthropic, 2024)	2.0	32.5	42.8	23.2	26.9	0.7	57.2	0.0	24.8	21.1	46.2	22.3	858.6
Open-Source Models													
Llama-4-Maverick-17B-128E (Singh, 2025)	24.5	32.5	41.0	25.0	29.1	4.5	47.6	4.6	15.2	29.6	30.8	23.8	845.2
Kimi-VL-A3B-Thinking (Team et al., 2025)	2.0	33.3	31.7	50.0	23.8	0.0	47.6	0.6	23.2	5.6	11.5	18.2	785.2
Qwen2.5-VL-72B-Instruct (Bai et al., 2025)	2.0	33.3	35.9	23.2	26.9	0.0	47.6	0.0	20.8	17.8	19.2	19.7	497.6
GLM-4.5V (Hong et al., 2025)	4.1	32.5	43.4	30.4	23.0	1.5	57.9	0.0	22.4	14.1	30.8	21.0	1391.1
Qwen2.5-VL-7B (Bai et al., 2025)	0.0	32.5	34.8	17.9	24.4	0.0	41.4	0.0	20.0	13.6	15.4	17.8	465.3
InternVL3-8B (Zhu et al., 2025)	0.0	29.3	35.9	23.2	21.3	0.7	40.0	0.0	19.2	14.6	15.4	17.4	473.5
DeepSeek-VL2-7B (Wu et al., 2024)	0.0	14.6	11.7	7.1	5.9	4.5	2.1	2.1	11.2	4.2	7.7	6.6	397.4
Pixtral-12B (Agrawal et al., 2024)	4.1	22.8	34.1	26.8	20.4	1.5	34.5	0.0	8.0	12.2	3.8	15.3	692.3
Idefics3-8B (Laurençon et al., 2024)	0.0	22.8	25.9	14.3	18.2	1.5	22.1	0.0	8.0	6.1	0.0	11.7	294.5
Llama3-8B (Dubey et al., 2024)	10.2	15.4	20.7	17.9	15.1	20.9	23.4	9.3	13.6	12.2	7.7	15.0	785.4
MathSpatial-Qwen2.5-VL-7B	4.1	40.7	46.2	30.4	27.7	0.7	46.2	0.0	21.6	17.4	26.9	21.9±0.8	351.9
MathSpatial-InternVL3-8B	2.0	50.4	36.2	33.9	22.7	11.9	56.6	3.9	23.2	15.5	15.4	22.6	318.3
MathSpatial-Llama3-8B	16.3	30.1	24.5	41.1	20.4	30.6	27.6	10.2	20.8	16.4	7.7	20.3	397.3
Human Performance													
Human	100.0	89.4	97.6	98.2	96.1	97.0	95.2	96.5	94.4	97.7	96.2	96.3	–

8B (Dubey et al., 2024). To demonstrate the training value of *MathSpatial-Corpus*, we fine-tune three open-source models (Qwen2.5-VL-7B, InternVL3-8B, Llama3-8B) on the corpus using a batch size of 128, learning rate 1×10^{-5} , and 5 epochs. To establish a rigorous human upper

bound, we recruited 80 students under closed-book conditions, with each problem answered by 20 independent annotators and results reported as micro-averaged accuracy. Each model is evaluated in two independent runs, and we expand MathSpatial-Qwen2.5-VL-7B, GPT-5.4, Gemini-3.1-pro, and

Claude Opus 4.6 to three runs; Table 2 reports mean $\pm$ standard deviation for these four models.

4.2 Benchmark Results

Table 2 presents the evaluation of 20+ MLLMs on *MathSpatial-Bench*. The results reveal a persistent human–model gap: humans exceed 95% accuracy across all categories, while most MLLMs remain below 60%, and even the strongest model falls $\sim$ 16 points short of human performance.

Reasoning models narrow but do not close the gap. Gemini-3-flash (Gemini Team, 2025b) achieves the highest accuracy (79.8%) and dominates nearly all subtypes, benefiting from its extended thinking capability (10,720 avg. tokens). Gemini-3.1-pro (Gemini Team, 2026) follows closely at $75.2\pm 0.6\%$. However, even these strongest reasoning models still fall 16–21 points below human performance (96.3%), indicating that mathematical spatial reasoning remains fundamentally challenging despite rapid model advances.

Standard models remain substantially weaker. Among standard (non-reasoning) closed-source models, GPT-5 (58.5%), GPT-5.4 ($54.2\pm 1.1\%$), and Claude Opus 4.6 ($47.8\pm 0.9\%$) lead, while the GPT-4 series (19.6%–22.6%) and older Claude family (21.5%–26.7%) lag significantly, scoring near zero on numerically demanding tasks such as GPC (0.0%–0.2%) and VT (0.0%–1.5%).

Open-source models lag substantially. Most open-source models cluster in the 15%–24% range. Even scaling up provides limited benefit: Qwen2.5-VL-72B (19.7%) only marginally improves over its 7B variant (17.8%). Multiple models score 0.0% on GPC, indicating a complete inability to perform geometric property calculations.

Difficulty varies sharply across reasoning categories. Holistic Recognition is the most accessible: top models reach 60%–83% on subtypes like IVI and 3VM. Generative Inference presents a mixed picture, with MVC well above chance for strong models but VT near zero for non-reasoning models. Abstract Deduction is the hardest category overall; while PRA is relatively tractable (Gemini-3-flash (Gemini Team, 2025b): 85.5%), GPC remains a near-universal failure point ($\leq 4\%$ for all open-source models), highlighting that multi-step geometric computation is a fundamental bottleneck.

Training on MathSpatial-Corpus demonstrates clear dataset value. Fine-tuning on *MathSpatial-Corpus* consistently lifts performance across all three backbone models. MathSpatial-InternVL3-

Table 3: Performance comparison on external OOD benchmarks.

Benchmark	Qwen2.5-VL-7B		MathSpatial (Ours)		Improvement	
	Acc $\uparrow$	Token $\downarrow$	Acc $\uparrow$	Token $\downarrow$	Acc $\uparrow$	Token $\downarrow$
SOLIDGEO	15.5	490.2	18.9	385.6	+3.4	-21.3%
GeoEval	17.6	395.4	19.4	367.5	+1.8	-7.1%
3DSRBench (Real)	48.4	752.9	49.7	643.2	+1.3	-14.6%

8B reaches 22.6% (vs. 17.4% base), achieving the second-highest 3VM score (50.4%) across all models. MathSpatial-Qwen2.5-VL-7B reaches $21.9\pm 0.8\%$ over three runs; its reference run scores 22.1%, with gains on IVI (+11.4%), PPCD (+11.5%), and 3VM (+8.2%). MathSpatial-Llama3-8B shows the largest VT improvement (20.9% $\rightarrow$ 30.6%). Notably, all fine-tuned models also reduce token consumption by 20%–30% (e.g., Qwen2.5-VL-7B: 465.3 $\rightarrow$ 351.9), demonstrating that the dataset’s structured annotations promote both accuracy and reasoning efficiency.

4.3 Generalization to External Benchmarks

To assess whether the improvements from *MathSpatial-Corpus* reflect genuine spatial reasoning acquisition rather than overfitting to the training distribution, we evaluate on three external out-of-distribution (OOD) benchmarks: SOLIDGEO (Wang et al., 2025a), GeoEval (Zhang et al., 2024), and the real-world subset of 3DSRBench (Ma et al., 2025).

As shown in Table 3, despite *MathSpatial-Bench* being designed to minimize perceptual dependencies, the reasoning skills acquired from our corpus transfer effectively to perception-heavy and visually complex domains. *MathSpatial-Qwen2.5-VL-7B* achieves consistent accuracy gains (+1.3% to +3.4%) while simultaneously reducing token usage by 7.1%–21.3%. These results confirm that the structured reasoning traces promote concise and generalizable reasoning patterns rather than dataset-specific memorization.

Furthermore, Figure 7 provides a qualitative illustration of why each atomic operation matters: removing any single operation (CORRELATE, CONSTRAIN, or INFER) from the structured trace leads to degraded or incorrect solutions, demonstrating that all three are essential for spatial reasoning.

4.4 Error Analysis

To better understand the limitations of current MLLMs on mathematical spatial reasoning, we conduct a fine-grained error analysis on

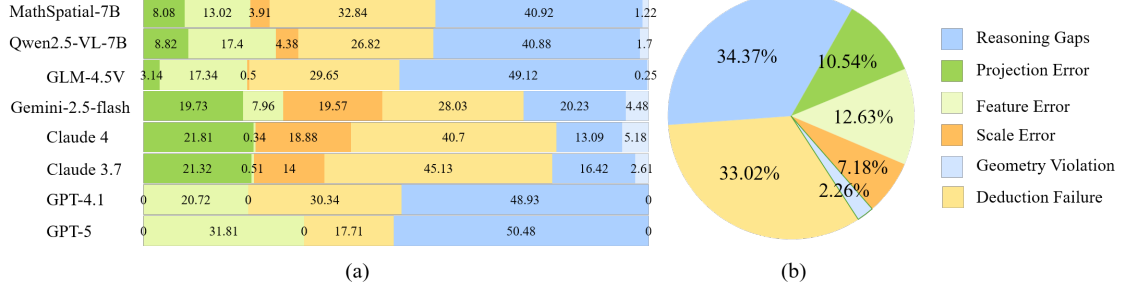


Figure 6: Fine-grained error analysis on *MathSpatial-Bench*. (a) Error frequency distribution for baselines across 6 subcategories. (b) Overall error rate breakdown by failure mode.

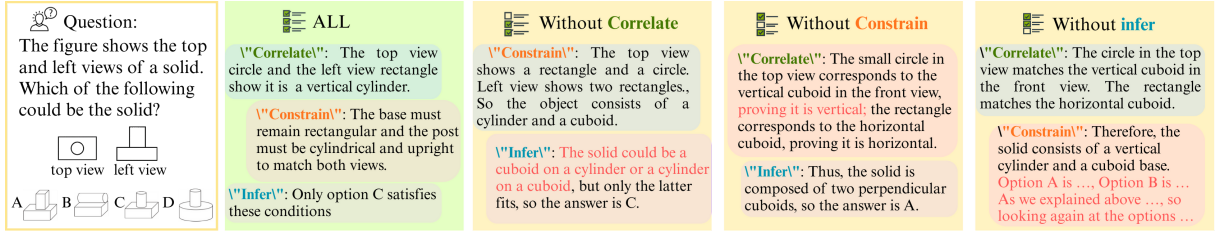


Figure 7: *MathSpatial-SRT* example: complete solution with all operations versus failures when removing Correlate, Constrain, or Infer.

MathSpatial-Bench. Errors are grouped into six categories: (1) Projection Error, where models misinterpret top, side, or front views; (2) Feature Error, such as omitting components or inventing edges; (3) Scale Error, failing to preserve relative sizes; (4) Geometry Violation, where outputs break orthographic or visibility rules; (5) Reasoning Gaps, incomplete or inconsistent chains of thought; and (6) Deduction Failure, where models cannot synthesize multiple cues into a final conclusion.

Figure 6(a) reveals distinct weakness profiles across model families: the GPT-5/4 series is dominated by reasoning gaps, Claude models frequently violate geometric rules, Gemini-2.5-flash suffers more from projection and scale errors, while open-source models like Qwen2.5-VL-7B exhibit primarily reasoning gaps but fewer feature errors. Notably, *MathSpatial-Qwen2.5-VL-7B* (fine-tuned on our corpus) reduces reasoning gaps compared to its base model, suggesting that the structured annotations in *MathSpatial-Corpus* help mitigate this dominant failure mode. As summarized in Figure 6(b), the majority of errors come from **reasoning gaps (34.4%)** and **geometry violations (33.0%)**, followed by projection (12.6%) and feature errors (10.5%). Scale (7.2%) and deduction failures (2.3%) are less frequent but still harmful.

These findings highlight two core challenges: (i) enforcing low-level geometric consistency across

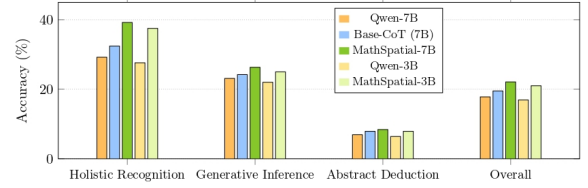


Figure 8: Ablation studies on *MathSpatial* across different training strategies (baseline, Base-CoT, SRT CoT) and model scales (3B, 7B).

views, and (ii) sustaining coherent multi-step reasoning chains. Both remain largely unsolved by most MLLMs, and the fine-grained error taxonomy provided by *MathSpatial-Bench* offers a diagnostic foundation for targeted improvements.

4.5 Ablation Studies

Ablation on Training Strategy. As shown in Figure 8(a), we compare three variants: vanilla baseline (Qwen2.5-VL-7B), free-form CoT baseline (Base-CoT)², and our structured reasoning traces (SRT CoT). SRT CoT improves accuracy from 17.8% to 22.1%, gaining +4.3% over baseline and +2.6% over Base-CoT. The largest improvement occurs in Holistic Recognition (+10.0%), with Generative Inference (+3.2%) and Abstract Deduction (+1.5%) also benefiting. Subtask analysis

²Same generation and training pipeline as *MathSpatial-SRT*, but uses free-form reasoning traces.

shows 3VM, IVI, and CR improve most, indicating that the CORRELATE/CONSTRAIN operations effectively reduce cross-view confusion and enforce geometric consistency. As further illustrated by Figure 7, removing any single atomic operation causes distinct degradation, confirming that each operation is indispensable.

Ablation on Model Scale. Figure 8(b) examines performance across model sizes. *MathSpatial-3B* outperforms its baseline (16.9% $\rightarrow$ 21.0%, +4.1%), mirroring the 7B trend: Holistic Recognition shows the largest improvement (+9.9%), with Generative Inference (+3.0%) and Abstract Deduction (+1.5%) also improving. Notable subtask gains appear in 3VM, IVI, CR, PRA, and PPCD, confirming the robustness of structured supervision across scales.

5 Conclusion

We presented *MathSpatial*, a large-scale, publicly available dataset for mathematical spatial reasoning in MLLMs. *MathSpatial* comprises *MathSpatial-Bench* (2K problems) for diagnostic evaluation across 3 categories and 11 subtypes, and *MathSpatial-Corpus* (8K problems) with verified solutions and structured reasoning annotations based on three atomic operations. Comprehensive benchmarking of leading MLLMs reveals a substantial human-model gap, with humans exceeding 95% accuracy while most models remain below 60%. Fine-grained error analysis identifies geometric rule violations (33.0%) and reasoning gaps (34.4%) as the dominant failure modes, offering actionable insights for future model development. Fine-tuning on *MathSpatial-Corpus* consistently improves both accuracy and efficiency across multiple model families, demonstrating the dataset’s practical value for advancing spatial reasoning capabilities. We hope *MathSpatial* will accelerate progress in multimodal mathematical spatial reasoning and related downstream applications.

Limitations

While our focus on educational geometry problems ensures strict control over perceptual complexity and enables the precise, high-quality annotations necessary for accurately isolating pure reasoning capabilities from basic visual perception, this targeted scope naturally differs from the unconstrained noise of real-world environments. Furthermore, although our framework systematically covers the static 2D and 3D geometric problems that constitute the foundational bedrock of spatial

cognition, it leaves dynamic, temporal-spatial scenarios for future exploration. Despite these deliberate scoping choices, *MathSpatial* establishes a robust and comprehensive framework for advancing spatial reasoning research in multimodal large language models and future embodied AI.

Acknowledgement

This work was supported by the New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123501), the Beijing Natural Science Foundation under Grant Z260008, and the National Natural Science Foundation of China under Grants U2441251 and 62276256. We extend our sincere thanks to the anonymous reviewers for their constructive suggestions. We also thank Meituan for providing hardware support for this work.

References

- Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, and 1 others. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.
- Anthropic. 2024. Claude 3.5 sonnet. <https://www.anthropic.com/news/claude-3-5-sonnet>.
- Anthropic. 2025a. Claude 3.7 sonnet. <https://www.anthropic.com/news/claude-3-7-sonnet>.
- Anthropic. 2025b. System card: Claude opus 4 & claude sonnet 4. <https://www.anthropic.com/claude-4-system-card>.
- Anthropic. 2026. Introducing claude opus 4.6. <https://www.anthropic.com/news/claude-opus-4-6>.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024a. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proc. CVPR*.
- Guhan Chen, Songtao Tian, Bohan Li, Hejin Wang, YeXin Xie, and Zixiong Yu. 2026. *Diag: Diagnostic iterative alignment and generation for data-efficient mathematical preference distillation. Preprint*, arXiv:2608.22806.
- Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Yan Feng, Peng Pei, Xunliang Cai, and Ruqi Huang. 2025. Think

- with 3d: Geometric imagination grounded spatial reasoning from limited views. *arXiv preprint arXiv:2510.18632*.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.
- Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, and 1 others. 2025. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In *Proc. ICCV*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*.
- Jie Feng, Jinwei Zeng, Qingyue Long, Hongyi Chen, Jie Zhao, Yanxin Xi, Zhilun Zhou, Yuan Yuan, Shengyuan Wang, Qingbin Zeng, and 1 others. 2025. A survey of large language model-powered spatial intelligence across scales: Advances in embodied agents, smart cities, and earth science. *arXiv preprint arXiv:2504.09848*.
- Google Gemini Team. 2025a. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). Technical report, DeepMind / Google.
- Google Gemini Team. 2025b. Gemini 3 flash-model card. <https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf>.
- Google Gemini Team. 2026. Gemini 3.1 pro. <https://deepmind.google/models/model-cards/gemini-3-1-pro/>.
- Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, and 1 others. 2025. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*.
- Yuchao Hou, Jiazhe Jiao, Jie Wang, Guangyin Jin, Zijian Zhang, Xiaoyu Xia, Zhiquan Liu, Minglu Li, and Youliang Tian. 2026a. Federated analytics assisted semantic alignment for secure and privacy-preserving image classification. *IEEE Transactions on Dependable and Secure Computing*.
- Yuchao Hou, Dongming Lv, Minghui Liwang, Jie Wang, Zijian Li, Xiaoyu Xia, Di Wu, Youliang Tian, Tony QS Quek, and Xianbin Wang. 2026b. Asyncfed-sar: towards communication-efficient asynchronous federated learning for sar image target recognition. *IEEE Transactions on Mobile Computing*.
- Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. 2025. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*.
- Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. 2025. Natural language understanding and inference with mllm in visual question answering: A survey. *ACM Computing Surveys*, 57(8).
- Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proc. ICML*.
- Linjie Li, Mahtab Bigverdi, Jiawei Gu, Zixian Ma, Yinyu Yang, Ziang Li, Yejin Choi, and Ranjay Krishna. 2025a. Unfolding spatial cognition: Evaluating multimodal models on visual simulations. *arXiv preprint arXiv:2506.04633*.
- Yuqi Li, Junhao Dong, Jiao Liu, Piotr Koniusz, Hansheng Zeng, Chuanguang Yang, Junming Liu, Yingli Tian, Tingwen Huang, and Hao Wu. 2026a. Evolving multimodal models for physical dynamics: A multi-objective neuroevolution approach. *IEEE Transactions on Evolutionary Computation*.
- Yuqi Li, Siwei Meng, Chuanguang Yang, Weilun Feng, Junming Liu, Zhulin An, Yikai Wang, and Yingli Tian. 2026b. A comprehensive survey of interaction techniques in 3d scene generation. *IJCAI*.
- Yuqi Li, Chuanguang Yang, Hansheng Zeng, Zeyu Dong, Zhulin An, Yongjun Xu, Yingli Tian, and Hao Wu. 2025b. Frequency-aligned knowledge distillation for lightweight spatiotemporal forecasting. In *Proc. ICCV*.
- Shuo Lu, Yanyin Chen, Wei Feng, Jiahao Fan, Fengheng Li, Zheng Zhang, Jingjing Lv, Junjie Shen, Ching Law, and Jian Liang. 2025a. Uni-layout: Integrating human feedback in unified layout generation and evaluation. In *Proc. ACM MM*.
- Shuo Lu, Haohan Wang, Wei Feng, Weizhen Wang, Shen Zhang, Yaoyu Li, Ao Ma, Zheng Zhang, Jingjing Lv, Junjie Shen, and 1 others. 2026a. One size, many fits: Aligning diverse group-wise click preferences in large-scale advertising image generation. In *Proc. CIKM*.
- Shuo Lu, Yingsheng Wang, Lijun Sheng, Lingxiao He, Aihua Zheng, and Jian Liang. 2025b. Out-of-distribution detection: A task-oriented survey of recent advances. *ACM Computing Surveys*, 58(2).

- Shuo Lu, YINUO Xu, Kecheng Yu, Siru Jiang, Yongcan Yu, Yubin Wang, Haitao Yang, Yuxiang Zhang, Bin Wang, Ran He, and 1 others. 2026b. Worldcoder-bench: Benchmarking physically grounded 3d world synthesis. *arXiv preprint arXiv:2606.01869*.
- Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 2025. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proc. CVPR*.
- OpenAI. 2024. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>.
- OpenAI. 2025. Gpt-5 system card. Technical report, OpenAI. Accessed: 2025-08-10.
- OpenAI. 2025. [Introducing gpt-4.1 in the api](#).
- OpenAI. 2026. [Introducing gpt-5.4](#).
- Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, and 1 others. 2025. Humanity’s last exam. *arXiv preprint arXiv:2501.14249*.
- Jun Rao, Zixiong Yu, Xuebo Liu, Guhan Chen, Jing Li, Hejin Wang, Jiansheng Wei, Xiaojun Meng, and Min Zhang. 2026. [Mining or synthesis? rethinking exploration efficiency in iterative alignment of mathematical reasoning](#). *Preprint*, arXiv:2602.05370.
- Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, and 1 others. 2024. Sat: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*.
- Sara Sarto, Marcella Cornia, Rita Cucchiara, and 1 others. 2025. Image captioning evaluation in the age of multimodal llms: Challenges and future perspectives. In *Proc. IJCAI*.
- Ajit Singh. 2025. [Meta llama 4: The future of multi-modal ai](#). Available at SSRN 5208228.
- Kexian Tang, Junyao Gao, Yanhong Zeng, Haodong Duan, Yanan Sun, Zhening Xing, Wenran Liu, Kaifeng Lyu, and Kai Chen. 2025. Lego-puzzles: How good are mllms at multi-step spatial reasoning? *arXiv preprint arXiv:2503.19990*.
- Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, and 1 others. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.
- Kexin Tian, Jingrui Mao, Yunlong Zhang, Jiwan Jiang, Yang Zhou, and Zhengzhong Tu. 2025. Nuscenes-spatialqa: A spatial understanding and reasoning benchmark for vision-language models in autonomous driving. *arXiv preprint arXiv:2504.03164*.
- Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. 2024. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. In *Proc. NeurIPS*.
- Peijie Wang, Chao Yang, Zhong-Zhi Li, Fei Yin, Dekang Ran, Mi Tian, Zhilong Ji, Jinfeng Bai, and Cheng-Lin Liu. 2025a. Solidgeo: Measuring multimodal spatial math reasoning in solid geometry. In *Proc. NeurIPS*.
- Siting Wang, Luoyang Sun, Cheng Deng, Kun Shao, Minnan Pei, Zheng Tian, Haifeng Zhang, and Jun Wang. 2025b. Spatialviz-bench: Automatically generated spatial visualization reasoning tasks for mllms. *arXiv preprint arXiv:2507.07610*.
- Yanbo Wang, Yongcan Yu, Jian Liang, and Ran He. 2025c. A comprehensive survey on trustworthiness in reasoning with large language models. *arXiv preprint arXiv:2509.03871*.
- Yi Wang, Lanling Zeng, Jiaqi Zhang, and Yang Yang. 2026. Zero-shot diffusive image restoration with consistency. *Signal Processing*, 248.
- Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. 2025. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*.
- Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.
- Guo Yang, Jiaqi Zhang, Yao Zhu, and Longze Fan. 2026a. Agent-sam-i2v: Self-correcting promptable video segmentation via agentic drift detection and multi-prompt fusion. In *Proc. ICIC*, volume 16662 of *Lecture Notes in Computer Science*. Springer.
- Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proc. CVPR*.
- Zhen Yang, Sizai Hou, Kaiwen Zheng, Yaofang Liu, Liang He, Yixuan Chen, and Kangning Cui. 2026b. [Target-aware calibration data selection for preserving uncertainty in quantized language models](#). *Preprint*, arXiv:2608.21019.
- Shaokai Ye, Haozhe Qi, Alexander Mathis, and Mackenzie W Mathis. 2025. Llavaction: evaluating and training multi-modal large language models for action recognition. *arXiv preprint arXiv:2503.18712*.
- Yongcan Yu, Lingxiao He, Shuo Lu, Lijun Sheng, YINUO Xu, Yanbo Wang, Kuangpu Guo, Jianjie Cheng, Meng Wang, Qianlong Xie, and 1 others. 2025. Reassessing the role of supervised fine-tuning: an empirical study in vlm reasoning. *arXiv preprint arXiv:2512.12690*.

Zixiong Yu, Jun Rao, Guhan Chen, Songtao Tian, Bohan Li, Jiansheng Wei, Min Zhang, and Xiaojun Meng. 2026. *Mathagent: Adversarial evolution of constraint graphs for mathematical reasoning data synthesis*. In *Proc. ACL Findings*, San Diego, California, United States. Association for Computational Linguistics.

Yifu Yuan, Haiqin Cui, Yibin Chen, Zibin Dong, Fei Ni, Longxin Kou, Jinyi Liu, Pengyi Li, Yan Zheng, and Jianye Hao. 2025. From seeing to doing: Bridging reasoning and decision for robotic manipulation. *arXiv preprint arXiv:2505.08548*.

Jirong Zha, Yuxuan Fan, Xiao Yang, Chen Gao, and Xinlei Chen. 2025. How to enable llm with 3d capacity? a survey of spatial reasoning in llm. In *Proc. IJCAI*.

Jiaqi Zhang, Zheng Pang, Rongrong Gao, Qiyuan Zhang, and Yang Yang. 2026. Local epistemic uncertainty guided active sampling for plug-and-play diffusive image restoration. *arXiv preprint arXiv:2608.06981*.

Jiaxin Zhang, Zhong-Zhi Li, Ming-Liang Zhang, Fei Yin, Cheng-Lin Liu, and Yashar Moshfeghi. 2024. Geoeval: benchmark for evaluating llms and multimodal models on geometry problem-solving. In *Proc. ACL*.

Yixiao Zhou, Dongzhou Cheng, Zhiliang Wu, Yi Yang, Yu Cheng, and Hehe Fan. 2026. One refiner to unlock them all: Inference-time reasoning elicitation via reinforcement query refinement. In *Proc. ACL*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Appendix

A.1 Annotation Guidance

To complement the main pipeline description (Figure 2), we provide detailed guidance on how annotation was conducted in practice. This section elaborates on the interfaces, annotator workflow, and validation mechanisms to ensure that the resulting dataset is both reliable and reproducible.

A.1.1 Data Collection and Curation

Source Diversity and Bias Mitigation. Problems were systematically sourced from multiple online repositories, including Baidu Wenku, Zujian, and provincial exam archives. This strategy ensures broad coverage across difficulty levels and problem styles, effectively mitigating stylistic and regional bias. To avoid subjective labeling bias, we strictly

include only problems with objective numeric or multiple-choice answers.

Initial Filtering. A total of 35,428 raw candidates were collected. Annotators systematically removed structurally incomplete pages, non-spatial questions (e.g., arithmetic, logic puzzles), and corrupted scans or duplicated PDF segments. After this rigorous filtering, 21,673 spatially valid items remained, corresponding to a 61.1% retention rate.

Metadata Recording. Each retained problem was tagged with rich metadata, including its origin, difficulty level (easy/medium/hard), and specific geometric topic (e.g., orthographic projection, 3D geometry, view-matching).

A.1.2 Preprocessing and Standardization

Unified Schema. Every item was standardized into a JSON structure: (images, questions, choices, answer, solution).

Duplicate Detection. We applied a rigorous multi-stage filtering pipeline combining hashing and semantic analysis. For image-level detection, we employed MD5 hashing for exact duplicates and GPT-4.1 vision similarity scoring for diagrams differing in resolution or scaling. For text-level detection, we utilized Sentence-BERT encodings. Let T_i and T_j be the textual content of two problems, and $\text{sim}(T_i, T_j)$ be their cosine similarity. The filtering action A is defined as:

$$A(T_i, T_j) = \begin{cases} \text{Discard} & \text{if } \text{sim}(T_i, T_j) \geq 0.90 \\ \text{Manual Review} & \text{if } 0.85 \leq \text{sim}(T_i, T_j) < 0.90 \\ \text{Retain} & \text{if } \text{sim}(T_i, T_j) < 0.85 \end{cases} \quad (1)$$

Following this automated screening, annotators reviewed approximately 1K cross-set candidate pairs, removing semantically equivalent statements or paraphrased variants. After preprocessing, approximately 11K unique and high-quality items remained.

Language Standardization. Original Chinese problems were translated into English using a commercial API. To ensure terminology consistency, annotators performed spot-checks on approximately 20% of the translated items.

A.1.3 Geometric Consistency Checking

Rule-Based Verification. Automated scripts enforced three core geometric constraints: dimensional correspondence across views (Front-Top-Side length/width/height alignment), correct usage of dashed/solid lines for hidden/visible edges, and orthographic projection compliance.

Human-in-the-Loop Review. Approximately $1K$ items were flagged by the validator as potentially inconsistent. Annotators inspected all flagged cases; roughly 60% were correctable and repaired, while the remaining $\sim 0.4K$ showed irreconcilable inconsistencies and were discarded after senior review.

Codified QA Rubric & Consensus. All surviving items were evaluated according to a standardized geometric QA rubric governing view correspondence logic and admissible projection conventions. Items where annotators could not reach consensus were strictly discarded.

A.1.4 Solution Verification

Cases With Official Solutions. For problems with authoritative solutions, we retained them directly after formatting verification to maintain accuracy.

Cases Without Solutions. For the subset of items lacking official solutions, we implemented a strict generation protocol. Six graduate annotators with backgrounds in geometry or engineering drawing independently solved assigned subsets. Each solution was subsequently cross-checked by two additional annotators, and only problems with unanimous agreement were accepted. The protocol required full justification of intermediate reasoning steps and the use of explicit spatial correspondences.

A.1.5 Final Dataset Output

Corpus Scale and Partitioning. The final dataset consists of over $10K$ fully curated problems. We release two subsets: *MathSpatial-Bench* ($2K$ items) for systematic evaluation and *MathSpatial-Corpus* ($8K$ items) for supervised training.

Unique Properties. MathSpatial combines a controlled focus on spatial reasoning (minimizing perception-heavy clutter) with authoritative solution annotations and operation-typed traces at scale.

A.2 Dataset

To provide a detailed understanding of **MathSpatial-Bench**, we first present its overall composition and subcategory definitions, followed by representative examples from each subcategory. Together, these provide both a systematic and an intuitive view of the benchmark.

A.2.1 Subcategory Descriptions

We organized the 2000 problems into 11 fine-grained subcategories, grouped under three reasoning levels: *Holistic Recognition*, *Generative*

Inference, and *Abstract Deduction*. Table 4 summarizes each subcategory with its name, brief description, problem count, and proportion, providing a structured overview of the dataset’s coverage and distribution.

A.2.2 Case Studies

To further illustrate the dataset, we provide one representative problem from each subcategory. As shown in Figure 9, the examples are arranged in three columns, corresponding to the three reasoning levels (*Holistic Recognition*, *Generative Inference*, *Abstract Deduction*), allowing readers to visually grasp the diversity of problem types and the nature of the reasoning challenges in each category.

A.3 More Details about Results

A.3.1 Comparison Results Analysis

In this section, we provide a detailed performance analysis of state-of-the-art models on MathSpatial-Bench, covering a diverse range of both closed-source and open-source systems. (1) Among closed-source models, the reasoning models Gemini-3-flash (Gemini Team, 2025b) (79.8%) and Gemini-3.1-pro (Gemini Team, 2026) ($75.2 \pm 0.6\%$) achieve the highest overall performance, followed by GPT-5 (58.5%) and GPT-5.4 ($54.2 \pm 1.1\%$). Claude Opus 4.6 and Sonnet 4.6 reach $47.8 \pm 0.9\%$ and 47.9% respectively, substantially outperforming the older Claude and GPT-4 series ($19\text{--}27\%$). Despite being the strongest model, Gemini-3-flash still falls ~ 16 points below human accuracy (96.3%). (2) Open-source models perform worse overall, with Qwen2.5-VL-7B, InternVL3-8B, and Llama3-8B scoring between $15\text{--}18\%$. Scaling to Qwen2.5-VL-72B yields modest improvements (19.7%). These results highlight both the difficulty of MathSpatial-Bench and the value of structured supervision in closing the gap between open-source and proprietary MLLMs.

A.3.2 Human Evaluation Protocol

We established the human baseline through a systematic evaluation protocol rather than ad-hoc sampling to ensure the statistical significance and reproducibility of the 96.3% score. The exact specifications of our study are as follows.

Participants and Demographics. We recruited $N_p = 80$ middle and high school students from standard academic tracks who had completed the corresponding mathematics curriculum but received no task-specific training. This setup ensures

Holistic Recognition

• **Question:** A table has several plates placed on it. From three different angles, the views appear as shown in the diagram. Determine the total number of plates on the table.

top view front view left view

A 10 B 11 C 12 D 13

Answer: C
Subcategory: C3M

Generative Inference

• **Question:** A geometric figure is composed of several identical small cubes. The shapes seen from the front and the top are as shown in the diagram. What is the maximum number of such small cubes needed to construct this geometric figure?

front view top view

A. 8 B. 9
C. 10 D. 11

Answer: D
Subcategory: MVC

Abstract Deduction

• **Question:** The three views of a solid are shown in the figure, where the front view is an equilateral triangle. What is the surface area of the circumscribed sphere of the solid?

front view left view top view

A. $(8/3)\pi$ B. $(16/3)\pi$ C. $(48/3)\pi$ D. $(64/3)\pi$

Answer: D
Subcategory: GPC

Holistic Recognition

• **Question:** As shown in the figure, the main view of a solid composed of several small cubes is?

A B C D

Answer: C
Subcategory: 3VM

Generative Inference

• **Question:** The solid shown in the figure is constructed from 5 identical small cubes. The shape seen from the left side is?

A B C D

Answer: D
Subcategory: VT

Abstract Deduction

• **Question:** The top view of a geometric figure constructed from five identical small cubes is shown in the diagram. The number of possible ways to arrange this geometric figure is?

A. 1 B. 2
C. 3 D. 4

Answer: D
Subcategory: FPBI

Holistic Recognition

• **Question:** As shown in the figure, which of the following horizontally placed solids does not have a rectangular left view?

A B C D

Answer: B
Subcategory: IVI

Generative Inference

• **Question:** As shown in the figure, the rim of the cup is parallel to the projection plane, and the direction of the projection line is indicated by the arrow. Its orthographic projection is?

A B C D

Answer: D
Subcategory: PRA

Abstract Deduction

• **Question:** As shown in the figure, triangle $\triangle ABC$ is translated along the direction of BC to the position of $\triangle DEF$. If $EC = 2BE = 2$, then the length of CF is?

A. 1 B. 2 C. 3 D. 4

Answer: D
Subcategory: SCP

Holistic Recognition

• **Question:** As shown in the figure are the front view and top view of a geometric solid. Then this geometric solid is?

front view top view

A. Triangular Prism
B. Cube
C. Triangular Pyramid
D. Rectangular Prism

Answer: A
Subcategory: CR

Abstract Deduction

• **Question:** As shown in the figure, a rectangle with a length of 2 and width of 1 moves horizontally from the left side of line L to the right side in the "posture" shown (the dashed line in the figure is the horizontal line). The distance of the translation is?

A. 1 B. 2
C. 3 D. $2\sqrt{2}$

Answer: C
Subcategory: PPCD

Figure 9: Example problems from each of the 11 subcategories in **MathSpatial-Bench**, covering the three reasoning levels: *Holistic Recognition*, *Generative Inference*, and *Abstract Deduction*.

Table 4: Organization of the *MathSpatial Benchmark* (2,000 problems). Each subcategory lists the number of problems (#) and proportion (%). The benchmark spans three levels: Holistic Recognition, Generative Inference, and Abstract Deduction.

Level	Subcategory	Brief Description	#	(%)
Holistic Recognition	C3M	Complex 3-view $\rightarrow$ 3D matching; select the unique 3D model consistent with three views (hard distractors: symmetry, hidden/occluded features).	49	2.5
	3VM	3D model $\rightarrow$ 3-view matching; given a 3D model, choose the valid set of orthographic projections.	123	6.2
	IVI	Incorrect view identification; detect the wrong (or only correct) view given a 3D object or a pair of correct views.	290	14.5
	CR	Component recognition; identify the correct Boolean composition of complex structures.	56	2.8
Generative Inference	MVC	Missing-view completion; infer a missing orthographic view from given ones under standard projection rules.	357	17.9
	VT	Viewpoint transformation; predict the target-perspective view (tests mental rotation and depth consistency).	134	6.7
	PRA	Projection-rule application; enforce dashed/solid line conventions and visibility across views.	145	7.3
Abstract Deduction	GPC	Geometric property calculation; compute volume, area, angle, centroid, etc., from multi-view or 3D cues.	482	24.1
	FPBI	Function & physical behavior inference; predict stability, support, rolling tendency, or interlock.	125	6.3
	SCP	Structural change prediction; reason about effects of edits (add/remove parts, cutouts, holes) on views or properties.	213	10.7
	PPCD	Path planning & collision detection; infer feasible paths and collision outcomes in constrained layouts.	26	1.3

that the human baseline reflects typical student-level spatial reasoning ability, rather than expert-level annotation or test familiarity.

Scale and Redundancy. To ensure statistical stability, we enforced a high redundancy coverage factor (C). Given the total number of problems in the benchmark $N_{\text{total}} = 2,000$, and each participant solving a subset $N_{\text{sub}} = 500$, the coverage is calculated as:

$$C = \frac{N_p \times N_{\text{sub}}}{N_{\text{total}}} = \frac{80 \times 500}{2,000} = 20\times \quad (2)$$

This $20\times$ redundancy minimizes sampling bias and provides stable, reproducible accuracy estimates.

Testing Conditions. All participants solved problems individually under strict closed-book conditions. The use of calculators, search engines, or geometry software was strictly prohibited; only pen and paper were permitted for auxiliary sketches. This ensures that the results reflect unaided human reasoning.

Accuracy Metric. The reported score is the micro-averaged accuracy over the total response set $\mathcal{R}$ ($|\mathcal{R}| = 40,000$). Let $\hat{y}_i$ be the participant’s answer for the i -th response and y_i^* be the ground truth. The human accuracy $\text{Acc}_{\text{human}}$ is defined as:

$$\text{Acc}_{\text{human}} = \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} \mathbb{I}(\hat{y}_i = y_i^*) = 96.3\% \quad (3)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function, equal to 1 if the condition holds and 0 otherwise.

A.4 More Details about MathSpatial-SRT

A.4.1 PROOFS FOR MATHSPATIAL-SRT

Proposition 1 (Normal-form coverage). For every task in MathSpatial-Bench (the 11 subcategories), there exists a finite sequence of operations over $\{\text{CORR}, \text{CONS}, \text{INFER}\}$ that solves the task; moreover, any valid solution can be transformed into an equivalent sequence in the *correlate* $\rightarrow$ *constrain* $\rightarrow$ *infer* normal form.

Proof.

(i) **Decomposition.** Each subcategory reduces to three stages: - identify correspondences across views (CORR); - enforce projection/geometric rules (CONS); - produce numeric or categorical outputs (INFER). For example, tasks in Holistic Recognition (C3M, 3VM, IVI, CR) require establishing cross-view feature identity and validating composition rules, while Generative Inference tasks (MVC, VT, PRA) require applying projection conventions, and Abstract Deduction tasks (GPC, FPBI, SCP, PPCD) demand property calculation or feasibility inference.

(ii) **Closure.** CORR only augments the correspondence set, while CONS only augments the rule

set. Both are monotonic updates, and intermediate queries can be treated as INFER operations that read from the accumulated state without invalidating earlier facts. Thus, any arbitrary solution sequence can be reordered into blocks of correlations, then constraints, followed by inference.

(iii) **Task coverage.** Each of the 11 subcategories can be mapped into this three-stage form, ensuring full expressivity.

Proposition 2 (Minimality). The primitive set $\{\text{CORR}, \text{CONS}, \text{INFER}\}$ is minimal: removing any one element strictly reduces expressivity.

Proof.

- Without CORR, cross-view identity tasks (e.g., C3M, 3VM, MVC) cannot be solved.
- Without CONS, tasks that require validating geometric or projection rules (e.g., PRA, CR, SCP) cannot be handled.
- Without INFER, tasks requiring numeric calculation or discrete decision outputs (e.g., GPC, FPBI, PPCD) cannot be completed.

Therefore, no primitive is redundant.

We clarify that these proofs apply within the formalized problem space of *MathSpatial*. The claim of minimality is further justified by *semantic orthogonality*: each operation targets a distinct, non-overlapping role—CORR for perceptual alignment, CONS for law application, and INFER for logical deduction. Merging any two would compromise the granularity required for precise error diagnosis (as evidenced in Error Analysis). Furthermore, the practical sufficiency of this set is empirically validated by our OOD generalization results (Section 4.3), which demonstrate that models trained on these primitives successfully transfer to complex, perception-heavy benchmarks (e.g., 3DSRBench-Real).

A.4.2 SRT Validation Workflow

Dual-Role Validation Workflow. To ensure reliability, GPT-4o reformats verified official solutions into our SRT schema (CORR/CONS/INFER) rather than generating solutions from scratch. A **Reviewer** checks each step for logical consistency, operation-type correctness, redundancy, and hallucination; a **Checker** then repairs the trace and enforces the $\text{CORR} \rightarrow \text{CONS} \rightarrow \text{INFER}$ flow. Unresolved structural conflicts or unstated assumptions are forwarded to human experts.

Validation Statistics. Approximately **90%** of traces pass the initial review, **7%** require auto-

matic repair, and **3%** receive human correction. A random-sample audit estimates the residual error rate below **1%** using a Clopper–Pearson interval with 95% confidence. Reviewer and Checker prompts are provided in Tables 5–6.

Table 5: Prompt for SRT Reviewer

SRT Reviewer Prompt
You are a Reviewer for a structured reasoning trace (SRT). Your task is to audit each step for correctness and consistency. Rules: • Step-by-step verification: check if each step is logically valid. • Operation-type correctness: CORRELATE / CONSTRAIN / INFER must be used appropriately. • Consistency: detect contradictions with the authoritative solution. • No hallucinations: no invented points, edges, or faces. • Conciseness: flag redundant or unnecessary steps. Inputs: Problem Statement, Diagram Description, Ground-Truth Answer, Official Solution, Candidate SRT. Output: A JSON report: • <code>is_valid</code>: bool • <code>issues</code>: list of {<code>step_index</code>, <code>issue_type</code>, <code>explanation</code>, <code>suggested_fix</code>}

Table 6: Prompt for SRT Checker Prompt.

SRT Checker Prompt
Instruction. You are a Checker and Rewriter. Your task is to repair the SRT based on the diagnostic report while preserving the MathSpatial-SRT schema. Task. • Fix all issues identified by the Reviewer. • Ensure strict adherence to CORRELATE / CONSTRAIN / INFER semantics. • Maintain consistency with the authoritative solution. • Produce a final answer matching the ground truth. Inputs. Problem statement, diagram, official solution, original SRT, Reviewer diagnostic report. Output. • <code>Corrected_SRT</code>: the repaired reasoning trace, • <code>Final_answer</code>: must match the ground truth, • <code>Self_check</code>: confirmation of consistency.

Additional validation note. The operation labels provide a local account of what each step does. CORRELATE links entities already present in the inputs, CONSTRAIN applies a stated spatial or geometric relation, and INFER derives the requested result from those facts. This separation allows the Reviewer to identify whether an error comes from alignment, rule application, or deduction. The Checker may simplify or relabel a trace, but preserves the supplied evidence, schema, and ground-truth answer; unresolved conflicts remain subject to human review.