

GGBench: A Geometric Generative Reasoning Benchmark for Unified Multimodal Models

Jingxuan Wei^{1,2,4,5}, Caijun Jia^{2,4*}, Xi Bai^{2,4*}, Xinglong Xu^{2,4*}, Siyuan Li³, Linzhuang Sun^{2,4},
Bihui Yu^{2,4†}, Conghui He³, Lijun Wu³, Cheng Tan^{3†}

¹Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences) ²Shenyang institute of computing technology, Chinese academy of sciences ³Shanghai Artificial Intelligence Laboratory ⁴University of Chinese Academy of Sciences ⁵Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science

weijingxuan20@mails.ucas.edu.cn, chengtan9907@gmail.com

Abstract

Unified Multimodal Models (UMMs) are redefining the landscape of artificial intelligence by coupling perception and generation across language, vision, and structured reasoning. Yet, despite their growing sophistication, a critical gap persists in evaluation: existing benchmarks largely measure discriminative understanding or unconstrained generation in isolation, overlooking the integrated generative reasoning required for genuine multimodal intelligence. To address this, we introduce GGBench, the benchmark explicitly designed to evaluate geometric generative reasoning—the ability of a model to understand, reason about, and construct a solution within a unified framework. Each instance in GGBench contains precisely aligned natural-language instructions, executable GeoGebra code, and rendered diagrams, enabling deterministic and interpretable verification of a model’s reasoning and constructive fidelity. The benchmark comprises 1,411 rigorously curated problems covering eight categories and multiple difficulty levels, resulting in over 7,000 aligned visualizations. We propose a comprehensive tri-modal evaluation protocol that jointly assesses textual planning quality, code executability, and geometric accuracy of generated diagrams through both automated and human-in-the-loop judging. Extensive experiments on both state-of-the-art UMMs and general Large Language Models (LLMs) reveal a large performance gap between end-to-end generation and reasoning-grounded construction. GGBench establishes a new standard for testing multimodal systems that must not only understand but also build, marking a crucial

step toward grounded, verifiable generative intelligence.

1. Introduction

Artificial intelligence research is undergoing a profound paradigm shift, evolving from the foundational era of passive perception to the more sophisticated stage of contextual understanding [2], and now culminating in the frontier of active generation. This evolutionary leap is epitomized by the recent advent of UMMs [72], such as the groundbreaking GPT-4o [19] and Nano Banana [11], which demonstrate an unprecedented ability to fluently process, reason across, and synthesize information from a diverse spectrum of modalities. These UMMs can not only understand complex visual and textual inputs, but also generate rich outputs; notably, advanced systems like Unified-IO 2 [26] can even produce images, audio, or actions in response to multimodal instructions. Despite these breakthroughs, existing evaluation benchmarks lag behind. They primarily assess discriminative reasoning such as selecting an answer or classifying a visual input, and thus fail to capture the generative dimension of intelligence [59]. This limitation is particularly acute in geometry, where genuine comprehension is inseparable from the ability to plan and build a diagram via multi-step spatial reasoning [54]. In practice, even strong multimodal systems still struggle to interpret intricate geometric setups and to generate constructions that are both consistent and satisfy all constraints [12, 45].

The trajectory of evaluation benchmarks over recent years reflects a push toward increasingly complex cognition, yet also reveals a gap in evaluating **generative reasoning**. Early benchmarks targeted single-modality reason-

*Equal contribution

†Corresponding author.



Question : Using the digits {0, 6, 7, 8, 9}, how many two-digit numbers with distinct digits can be formed?



Solution : A two-digit number cannot begin with 0, so the tens place has 4 admissible choices {6, 7, 8, 9}. For each choice, the units place can be any of the remaining 4 digits (including 0). Hence the count is $(4 \times 4 = 16)$.
Equivalent argument: Total permutations of two distinct digits from five is $(P(5,2)=20)$. Exclude the 4 invalid numbers with leading zero $(0x, (x \in \{6, 7, 8, 9\}))$: $(20 - 4 = 16)$.
Answer: 16.

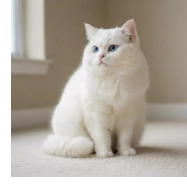
(a) Understanding Only



Question: Chubby white cat with blue eyes, sitting, photorealistic, neutral background.



Okay, here's that image for you:



(b) Generation Only



Question: Draw a regular pentagon inscribed in a circle.



Step 1: Construct a radius OP passing through the center O and perpendicular to OA

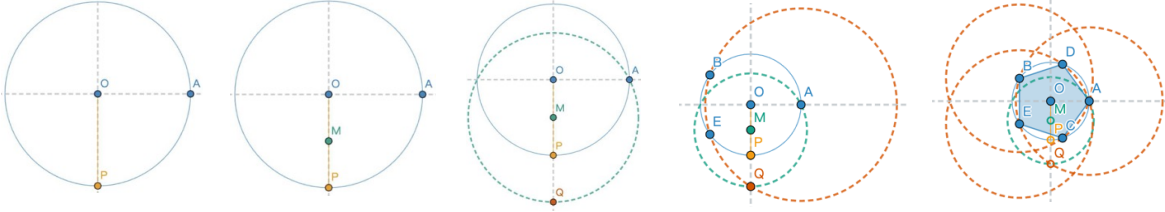
Construction:
1. Construct the line OA .
2. Construct a line l passing through point O and perpendicular to the line OA .
3. Line l intersects circle c at two points; select either intersection as point P .

Step 4: With center A and radius AQ , mark points B and E on circle c

Construction:
1. With point A as the center and the length of the line segment AQ as the radius, construct an auxiliary circle c_A .
2. Circle c_A intersects the original circle c at two points. These are two additional vertices of the regular pentagon, B and E .

Step 5: Sequentially mark the remaining vertices C, D and connect to form the regular pentagon

Construction:
1. With B as the center and a radius of AQ , draw a circle. This circle intersects circle c at A and another point, C .
2. With E as the center and a radius of AQ , draw a circle. This circle intersects circle c at A and another point, D .
3. Connect points A, B, C, D and E sequentially.



(c) GGBench: Understanding & Generation

Figure 1. GGBench’s step-by-step evaluation. Beyond traditional text-image pairs, GGBench provides executable code for each construction step, allowing for precise and automated verification.

ing or generation, primarily for LLMs in purely textual domains. For example, GSM8K [8] and MATH [16] established rigorous tests for mathematical problem-solving in text, requiring models to parse word problems and generate final answers through a step-by-step reasoning chain. Techniques such as chain-of-thought prompting greatly improved performance on these datasets [53]. Building on this progress, the community introduced multimodal reasoning challenges that incorporate visual context. Benchmarks like ScienceQA [27] require models to combine text understanding with image comprehension [44], and the recent Math-Vista suite [28] amalgamates 28 math and visual datasets to evaluate mathematical reasoning in visual contexts. These tasks, aimed at Multimodal LLMs (MLLMs), represent a critical step toward contextual understanding – models must interpret a question against an image or figure – but they remain largely discriminative. The model is still to answer or classify based on given content, rather than to create new content. The latest multimodal benchmarks have started to probe both understanding and generation capabilities, but often in a decoupled fashion. Efforts like MME [73], MM-Vet [66], and MMBench [25] evaluate a broad range of per-

ceptual and cognitive tasks to score models on each ability individually. Critically, however, even these comprehensive benchmarks treat understanding and generation as separate modules. This fragmentation means we lack evaluation of the integrated cognitive process – the scenario where a system must simultaneously understand, reason, and generate a complex outcome. We argue that geometric construction offers an ideal setting to assess this integrated ability—what we term **generative reasoning**. A capable model must:

- Parse natural-language instructions with domain-specific constraints;
- Formulate a multi-step plan grounded in formal geometric principles;
- Generate precise figures that satisfy all constraints.

Crucially, success is directly verifiable: constructions admit objective checks over objects and relations rather than subjective judgments alone.

To close this gap, we present GGBench, a benchmark purpose-built for geometric generative reasoning. As shown in Figure 1, each problem provides tri-modal supervision—aligned text with step-wise reasoning, executable GeoGebra code, and rendered diagrams—enabling determinis-

tic, step-by-step verification. GGBench supports both end-to-end UMM evaluation and modular diagnosis of planning and code generation. Our evaluation protocol can jointly measure (1) textual plan quality, (2) code executability and geometric equivalence, and (3) visual fidelity, combining both automated metrics with human review. As extensive experiments show, state-of-the-art UMMs exhibit a pronounced gap between answering about geometry and constructing geometry, underscoring the need for verifiable, construction-centric evaluation. GGBench establishes this rigor by turning “understand and reason” into “understand, reason, and build” manner.

2. Related Work

2.1. Mathematical Reasoning Benchmarks

Early benchmarks such as GeoQA [5] provides valuable testbeds but narrowly targets plane figures and evaluates models on answer accuracy. Several recent benchmarks like MATH-Vision [47], MathVista [28], and MathVerse [71] have broadened the scope of evaluation by combining visual and textual information. Specialized datasets like PolyMath [13] evaluates general cognitive reasoning in multimodal settings, requiring models to combine visual pattern recognition with mathematical logic. MathScape [75] uses real-world photographs for math questions. GeoEval [70] compiled a comprehensive suite of geometry question. SolidGeo [51] focuses on solid/3D geometry, whereas VisAidMath [30] evaluates the use of visual aids such as drawing auxiliary lines in geometry. In a different vein, MMSciBench [64] evaluates multimodal scientific problems, covering both math and physics questions in textual and text+image formats. VideoMathQA [37] assesses long-horizon mathematical reasoning in educational videos, and NewtonBench [74] challenges LLM-based agents to discover scientific laws through interactive experimentation. MathOPEval [21] requires models to perform fine-grained visual operations by generating executable code like constructing, modifying, or annotating a diagram. DynaMath [77] introduces dynamic variations of visual math problems to test robustness. MaRVL-QA [35] uses unlabeled function plots and geometry transformations as inputs that minimizes textual cues. MM-MATH [43] exemplifies this shift by augmenting outcome evaluation with process evaluation. We-Math [36] draws inspiration from human problem-solving principles and decomposes problem into sub-problems aligned with knowledge concepts. Math2Visual [46] introduces the task of generating pedagogically useful diagrams for math word problems. Most recently, Geoint-R1 [54] integrates auxiliary element construction with formal verification, and MathCanvas [41] enabled a unified multimodal model [10] to learn “when and how” to draw as part of its reasoning process.

2.2. The Rise of Unified Multimodal Models

UMMs signifies a pivotal architectural shift, not merely in the unification of modalities, but more fundamentally in the unification of understanding and generation capabilities within a single framework [72]. Recent examples of UMMs include GPT-4o [19] and Gemini 2.5 [9], as well as powerful open-source alternatives such as Janus series [6, 31, 55], Qwen-VL [4, 49], and OmniBridge [57], are all designed to seamlessly transition between comprehension and synthesis. Other efforts include Apple’s MM1 [32] and Google’s Nano Banana (Gemini 2.5 Flash Image) [11]. There are also hybrid systems like MM-ReAct [63], which augment an LLM with external visual tool interfaces rather than a single unified network. On the one hand, the understanding capabilities of these models have been rigorously validated across a vast landscape of benchmarks. They have achieved exceptional performance on testbeds like MMMU [67] for expert-level knowledge, MME-Unify [60] for holistic perception and cognition, and MathVista [28] for visual mathematical reasoning. Models explicitly designed for deep comprehension, such as Query-Kontext [42] and MAGUS [20], further demonstrate this proficiency. On the other hand, the generative prowess of UMMs is rapidly advancing, moving far beyond simple image captioning. Models like Bagel [10] and its successor Hyper-Bagel [29] showcase sophisticated generative abilities honed by novel training techniques. Application-specific models like ChartSketcher [18] can synthesize data visualizations from text, while generative models like Lingshu [61], VARGPT [76], and HaploOmni [58] exhibit advanced faculties for creating complicated and context-aware multimodal content. The development of sophisticated evaluation frameworks like UniEval [22] and reward models such as Skywork-VL Reward [52] reflects a growing effort to quantify and guide complex generative behaviors.

2.3. Code-based Evaluation for Reasoning Tasks

To enable rigorous evaluation, a growing body of work adopts code as a structured mechanism for assessment, leveraging the verifiable nature of code over free-form natural language explanations. MathCoder-VL [48] employs image-to-code supervision to ensure precise visual-textual alignment, while MATP-BENCH [15] integrates formal theorem proving in Lean, Coq, and Isabelle for verifiable multimodal logic. VeriEquivBench [68] proposes an equivalence-based validation metric for verifying consistency between generated code and specifications. InternLM-Math [65] unifies solving and verifying under a single framework with interleaved reasoning and code execution. DeepMath-103K [14] provides a large, contamination-free dataset with verifiable answers for reinforcement learning with rule-based rewards. MathQ-Verify [39] targets problem validity itself through a struc-

Table 1. Distribution of reasoning categories in GGBench. Each problem may involve multiple categories.

Category	Count
Basic Constructions	1,063
Circle Properties & Constructions	931
Geometric Transformations	376
Triangle Properties & Constructions	280
Applications of Geometric Theorems	218
Polygon Properties & Constructions	107
Measurement & Ratios	92
Locus Constructions	30
Total category tags	3,097

tured verification pipeline. CMMaTH [23] introduces a large Chinese multimodal math benchmark and an open-source automatic grader, while MARIO Eval [69] standardizes automated evaluation across mathematical datasets. U-MATH [7] assess university-level reasoning and meta-evaluate LLM judges. In parallel, QuesCo [33] captures holistic mathematical intent through contrastive pretraining. While recent multimodal datasets [28, 36, 40, 47, 50, 75] have advanced the field, they seldom enforce a verifiable alignment between the natural language reasoning steps.

3. The GGBench Benchmark

3.1. Overview

We introduce GGBench, a geometric generative reasoning benchmark designed to evaluate UMMs through end-to-end construction tasks. Unlike conventional discriminative benchmarks that evaluate models solely on final answer selection or numerical computation, GGBench operationalizes the assessment of whether a system can genuinely construct a geometrically valid solution.

A defining feature of GGBench is its tri-modal data structure, where each sample comprises precisely aligned text, code, and image components. It enables a uniquely versatile and diagnostic evaluation framework. For instance, the complete, end-to-end task of generating a final image from a textual prompt serves as the primary test for the integrated generative reasoning of UMMs. Concurrently, the benchmark can be decoupled to probe specific capabilities: the natural language reasoning steps (text) and their translation into executable programs (code) can be used to assess the logical planning and code generation faculties of LLMs. The provided GeoGebra code acts as an unambiguous ground truth, offering a deterministic method for verifying the geometric correctness of any generated figure. An overview of the dataset construction pipeline is shown in Figure 2, which defines the stage annotations (a)–(f) referenced throughout the paper.

3.2. Dataset construction

Data collection and candidate pooling We manually search the web for public geometry problems and assemble a pool that spans classical constructions and contest-style tasks (Figure 2, stage (a)). To scale selection beyond purely manual effort, we use an LLM for assisted tagging and filtering (LLM-i in the figure) and then conduct human checking to retain items that are suitable for construction—i.e., problems with unambiguous geometric dependencies and diagrammatically actionable conditions (Figure 2, stage (b)). This stage emphasizes diversity in objects and relations (e.g., circles, tangents, incenters, parallels) while discarding under-specified prompts.

Prompt design and problem adaptation To elicit end-to-end construction in GeoGebra only, we design a *composite* prompt that pairs a textual specification with exemplar GGB code (Figure 2, stage (c)). In parallel, we prepare a reformulation prompt family that converts the selected problems into construction-oriented statements with explicit auxiliary objects and ordered dependencies (Figure 2, stage (d)). In practice, GPT-5 [34] serves as the primary authoring model during adaptation (LLM-ii), while GGB-oriented prompt templates are informed by prior engineering with Gemini 2.5 Pro [9] to improve syntactic fidelity. The outcome is a set of adapted problems whose reasoning steps align one-to-one with executable GGB operations.

Solution generation and visualization Given an adapted problem and its data-generation prompts, GPT-5 [34] (LLM-iii) produces synchronized outputs: (i) stepwise reasoning text, (ii) executable GGB code, and (iii) a rendered diagram obtained by running the code (Figure 2, stage (e)). This stage focuses on faithful construction: steps materialize auxiliary objects (e.g., angle bisectors, tangents, parallels), code reflects those steps, and the rendered figure visualizes the target configuration. At this point we obtain roughly $\sim 10,000$ draft instances prior to filtering.

Automated screening and expert finalization The final stage involved a rigorous two-tier filtering process: (i) We perform LLM-based quality control (LLM-iv) along three axes—*code executability*, *logical consistency of the construction process*, and *diagram correctness* to remove low-quality drafts (Figure 2, stage (f)); (ii) All surviving instances underwent a review by domain experts who verified geometric correctness, the sufficiency of the construction, and the precise consistency across modalities. After this two-tier filtering, GGBench retains 1,411 high-quality items with tightly aligned text–code–image triplets. A detailed statistical analysis of the dataset is provided as below.

3.3. Dataset Analysis

Category Distribution and Difficulty GGBench covers eight major categories of geometric reasoning, as detailed in Table 1. Each problem may involve multiple skills, result-

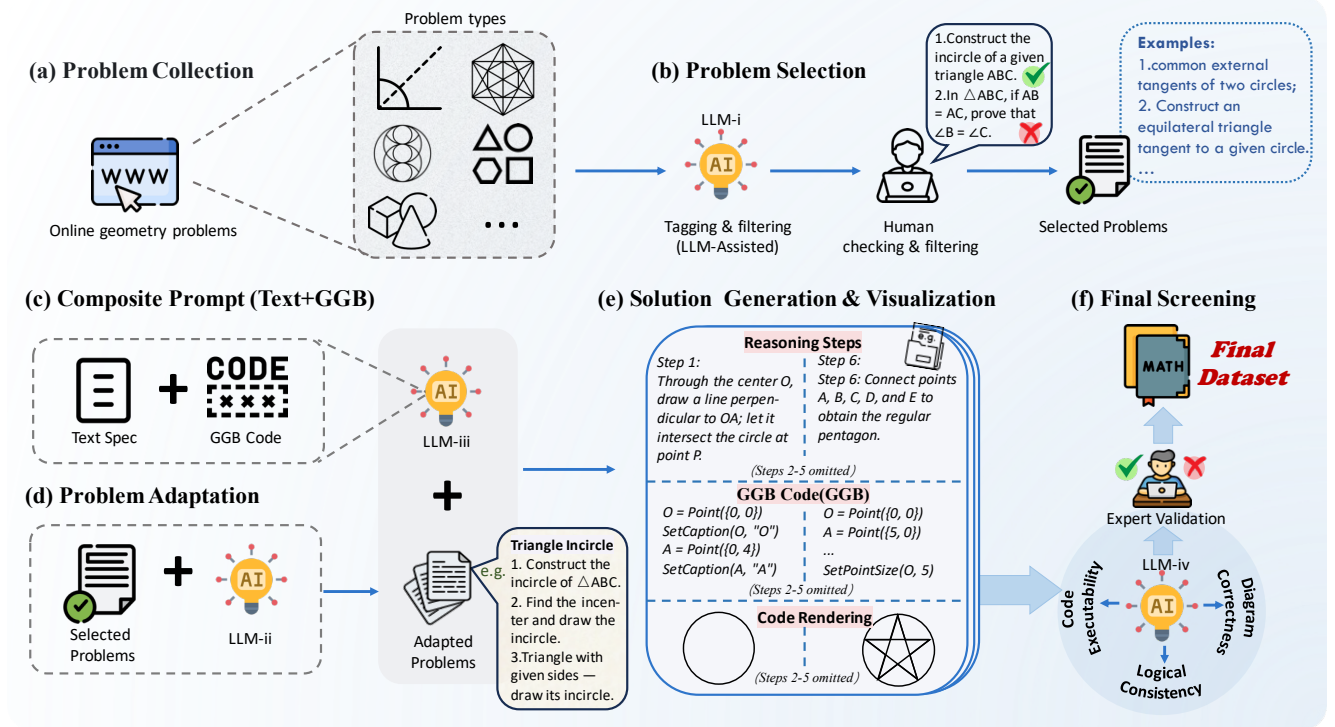


Figure 2. Overview of the GGBench data construction pipeline. Prompts (LLM i-iv) are detailed in Appendix A

ing in 3,097 category annotations. As shown in Figure 3, *Basic Constructions* and *Circle Properties & Constructions* are predominant across all difficulty levels, while advanced reasoning types—such as *Applications of Geometric Theorems* and *Measurement & Ratios*—become more frequent in the hard subset, ensuring balanced coverage from elementary to advanced geometric reasoning.

Dataset Composition and Scale GGBench comprises 1,411 GeoGebra-based construction problems that integrate reasoning, code, and visual outputs. Table 2 summarizes overall dataset statistics, including problem types, difficulty levels, image counts, and question lengths. Most problems belong to the medium difficulty (57.83%), balancing solvability and reasoning complexity. Straightedge-and-compass constructions dominate (56.6%), followed by geometric transformation and analytic constructions. Each problem includes between 3 and 7 diagrams, averaging 5.08 images per problem, providing dense visual evidence for evaluating generative reasoning.

Token Length and Reasoning Path As shown in Table 2, question length ranges from 68 to 483 tokens, averaging 189.83 tokens. It reflects the multi-step reasoning nature of GGBench, requiring the model to plan, describe, and validate geometric constructions. On average, each problem contains 5.08 images, and carries 2.20 category tags, indicating that GGBench is both visually dense and multi-skill by design. The five-image setting is the most common, providing ample visual evidence for stepwise con-

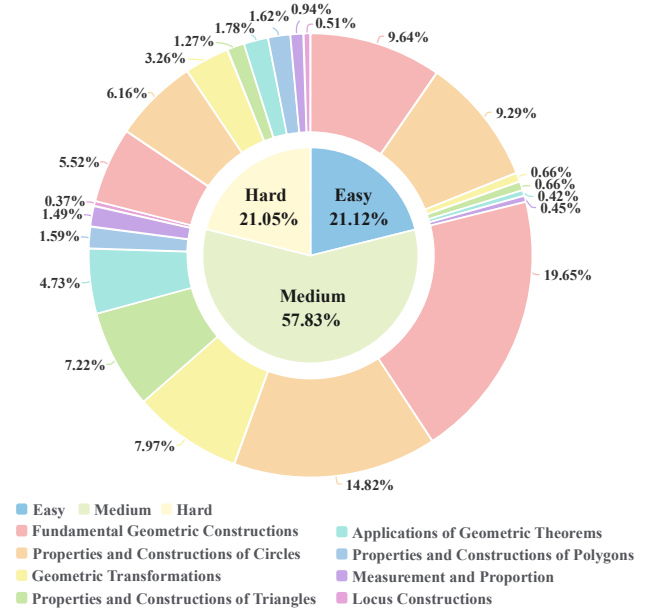


Figure 3. Difficulty distribution and category composition in GGBench. The inner ring shows the proportion of difficulty levels (Easy/Medium/Hard), while the outer ring presents category shares within each difficulty band.

struction. Figure 3 further shows a progressive shift toward higher-order skills (e.g., *Applications of Geometric Theorems*, *Measurement & Ratios*) in the hard subset.

Table 2. Corpus-level statistics of GGBench.

Problem types	
Straightedge-and-compass construction	798
Geometric transformation construction	426
Analytic construction	187
Difficulty levels	
Easy	298
Medium	816
Hard	297
Images per problem	
3 images	7
4 images	227
5 images	860
6 images	283
7 images	34
Total images	7,165
Question length (tokens)	
Minimum	68
Maximum	483
Average	189.83

3.4. Comparison with Existing Benchmarks

Table 3 situates GGBench among representative multi-modal math benchmarks along two groups of axes: (i) *capability coverage*—the proportion of samples that exercise *understanding*, *generation*, and *multi-step* reasoning; and (ii) *modality supervision*—whether problems provide aligned *text*, *images*, and *code*. We normalize modality coverage as a binary indicator in the table (✓/✗) for readability.

Generative construction vs. answer selection Most benchmarks [28, 75] emphasize recognition or answer selection. Even when generation is present [71], outputs are rarely required to be *constructive artifacts* that satisfy constraints. By contrast, every GGBench instance requires producing diagrams with formal constraints, shifting evaluation from choosing an answer to *constructing evidence*.

Tri-modal alignment and verifiability A central gap in prior work is the absence of code supervision. It is difficult to verify whether a model’s reasoning aligns with the final diagram. GGBench provides 100% alignment between *text* (stepwise plan), *code* (executable construction), and *image* (rendered diagram), enabling process-level judging, program executability checks, and geometry-aware verification within a single benchmark.

Multi-step supervision While some datasets include multi-step elements [43, 71], they do not couple each step with executable constructions, limiting verifiability. GGBench enforces multi-step reasoning at 100% and ties steps to code, ensuring that auxiliary objects and dependencies are operationalized rather than narrated.

Takeaway Together, these dimensions position GGBench as a construction-centric benchmark that unifies under-

standing, reasoning, and generation under a verifiable framework. It complements prior understanding-oriented datasets, offering a rigorous testbed for the next generation of unified multimodal reasoning models.

4. Experiments

We evaluate models under two complementary tracks that correspond to distinct architectural paradigms: (A) end-to-end UMMs that directly generate diagram images from natural-language prompts, and (B) LLMs/LRMs that first produce explicit, executable construction code which is then rendered into diagrams. This dual-track evaluation allows us to quantify and analyze the gap between UMMs’ immediate visual generation and the geometrically grounded, code-driven constructions produced by LLMs/LRMs. Detailed experimental setups are in the Appendix B.1.

4.1. Metrics

To evaluate geometric generative reasoning, We adopt a four-stage protocol: (1) *Planning*, (2) *Middle Process*, (3) *Final Result*, and (4) *Overall Scores*. All automatic scores are produced by a VLM GPT-4o using fixed prompts.

(1) Planning (VLM-T) measures the model’s ability to formulate a coherent step-by-step plan in natural language before any code or drawing is produced. Given the textual reasoning output, the VLM judge scores each response along three dimensions: (i) Logical coherence, (ii) Step completeness, and (iii) Geometric correctness. Each criterion is rated on a 1–5 scale and rescaled to [0,100]. Details of the judging rubric and prompt design appear in Appendix B.2.

(2) Middle process (VLM-I-Mid) To evaluate intermediate reasoning consistency, we concatenate all intermediate construction images into a chronological panel and feed it to the same VLM judge. The model assesses: (i) Step accuracy and (ii) Process consistency. Implementation and prompt details are provided in Appendix B.4.

(3) Final result (VLM-I-Res) assesses the geometric correctness of the final diagram relative to the reference solution. In addition, we report pixel-level metrics—LPIPS, PSNR, and SSIM. Details are in Appendix B.3.

(4) Overall scores The overall VLM score (VLM-I) is computed as the mean of the intermediate VLM-I-Mid and final scores VLM-I-Res. Human raters follow the same rubric as the VLM judge; inter-rater consistency and calibration procedures are detailed in Appendix B.6. We observe a strong Pearson correlation ($r = 0.9295$) between VLM and human scores, validating the high reliability of our automated evaluation for this task.

4.2. Main Results

Table 4 summarizes the overall performance across both tracks. End-to-end UMMs remain significantly behind code-driven models in nearly all evaluation dimensions.

Table 3. Comparison with existing multimodal mathematical benchmarks. Percentages under *Understanding/Generation/Multi-step* indicate the share of samples exercising each capability. *Text/Image/Code* report modality support (%).

Benchmark	Understanding	Generation	Multi-step	Text	Image	Code
MathVista [28]	49.9	34.8	21.5	✓	✓	✗
MathVerse [71]	61.3	52.7	45.2	✓	✓	✗
MM-MATH [43]	31.0	26.5	82.0	✓	✓	✗
MathScape [75]	44.2	39.8	18.5	✓	✓	✗
GeoEval [70]	46.0	25.0	33.3	✓	✓	✗
WE-MATH [36]	58.7	51.9	42.6	✓	✓	✗
MMSciBench [64]	55.4	32.8	40.2	✓	✓	✗
MATH2VISUAL [46]	48.0	37.0	✗	✓	✓	✗
PolyMath [13]	54.6	✗	✗	✓	✗	✗
SOLIDGEO [51]	49.5	18.7	6.7	✓	✓	✗
GGBench	100.0	100.0	100.0	✓	✓	✓

Table 4. Main results on GGBench. Higher is better (↑) except for LPIPS (↓).

Model	Planning	Middle Process	Final Result				Overall Scores	
	VLM-T ↑	VLM-I-Mid ↑	VLM-I-Res ↑	LPIPS × 10 ⁻² ↓	PSNR ↑	SSIM × 10 ⁻² ↑	VLM-I ↑	Human ↑
<i>End-to-end UMMs</i>								
Qwen-Image [56]	-	-	22.75	56.39	58.23	48.06	22.75	25.56
Seedream 4.0 [38]	-	-	24.45	51.06	59.44	56.44	24.45	37.56
Janus [55]	33.85	21.69	19.76	57.74	57.76	60.97	20.73	19.46
BAGEL [10]	23.07	21.84	19.99	57.07	61.78	58.82	20.91	20.12
Nano Banana [9]	58.54	44.83	22.81	51.85	64.53	59.51	33.82	45.75
<i>LLMs/LRMs</i>								
GPT-4o [19]	59.73	26.19	2.66	95.43	5.45	5.69	14.43	23.04
GLM-4.5V [17]	53.32	25.63	5.02	52.91	12.19	12.94	15.33	30.14
Qwen3-14B [62]	58.65	39.30	12.97	78.81	23.92	24.81	26.13	38.23
Gemini 2.5 Pro [9]	38.50	37.41	15.80	68.39	37.17	39.73	26.61	44.68
DeepSeek-R1 [24]	61.16	62.42	20.48	66.06	37.94	37.59	41.45	49.55
GPT-4 [1]	55.66	50.99	15.10	67.35	35.26	38.31	33.04	55.99
Qwen3-VL [62]	56.40	49.55	23.94	39.40	52.33	58.71	36.74	66.77
DeepSeek-V3.1 [24]	60.24	73.13	26.41	57.21	48.33	50.12	49.77	68.12
Claude Sonnet 4.5 [3]	61.19	77.92	30.29	52.22	51.74	50.52	54.11	72.12
GPT-5 [34]	62.01	76.79	37.36	49.65	54.80	59.49	57.08	83.06

Even the strongest UMM, *Nano Banana*, ranks only in the middle tier of code-based systems, illustrating that direct visual generation still struggles to enforce geometric constraints. Additional code-level metrics of LLMs/LRMs are reported in Table 5 of the Appendix.

The planning stage (VLM-T) evaluates a model’s capacity to reason step-by-step before any code or drawing is produced. Interestingly, *Nano Banana* achieves planning scores comparable to several LLMs/LRMs. However, models such as *GPT-4o* and *GLM-4.5V*, despite generating coherent plans, often fail to produce executable code, limiting their downstream visual accuracy. This gap demonstrates that high-level reasoning alone is insufficient without grounding in executable geometry.

The final-result scores (VLM-I-Res) show that code-driven models consistently satisfy geometric constraints

more precisely. Explicit reasoning followed by executable construction substantially outperforms direct visual generation. While good planning does not always guarantee perfect constructions, poor planning almost invariably leads to invalid diagrams—emphasizing the necessity of integrated reasoning-generation pipelines. Moreover, high PSNR or SSIM does not necessarily imply geometric correctness: pixel-level similarity can conceal structural errors such as misplaced intersections or missing tangencies. This observation reinforces the importance of reasoning-grounded evaluation over purely perceptual metrics.

4.3. Performance by Construction Category

Figure 4 presents VLM-I scores across eight construction categories. GPT-5 [34] consistently ranks highest, reflecting its strong capacity for both symbolic execution and

Locus Construction	57.333	62.000	49.500	47.833	41.167	39.000	36.333	35.333	17.500	22.833	21.333	21.333	20.500	26.333	21.833
Measurement & Ratios	65.978	59.402	52.989	36.957	34.511	31.685	32.120	35.543	27.228	22.446	22.826	22.609	22.391	14.565	17.337
Polygon Properties & Constructions	65.888	59.953	64.579	45.467	39.393	35.654	36.729	33.037	33.364	30.794	25.607	23.925	20.794	15.561	15.000
Application of Theorems	65.550	66.720	56.284	49.954	41.560	39.312	39.289	34.335	29.633	29.151	23.578	22.569	22.317	19.656	15.000
Triangle Properties & Constructions	60.518	58.446	52.661	40.982	39.089	35.393	34.875	40.661	32.000	28.268	23.786	22.143	24.250	18.268	15.750
Geometric Transformations	67.620	67.327	53.936	45.319	40.785	37.580	35.319	38.590	35.053	26.955	24.787	22.181	22.553	19.455	16.662
Circle Properties & Constructions	66.460	71.631	64.503	56.433	47.037	44.898	43.914	36.941	33.866	36.160	24.898	22.738	21.048	21.449	21.963
Basic Constructions	63.811	65.211	64.887	53.835	46.687	44.323	43.351	38.741	28.299	33.938	24.173	22.989	22.397	20.742	21.579
	Claude Sonnet	GPT-5	DeepSeek -V3.1	DeepSeek -R1	Janus	Qwen3 -VL	GPT-4	Nano Banana	Gemini 2.5 pro	Qwen3 -14B	Seedream 4.0	Qwen -Image	BAGEL	GLM -4.5V	GPT-4o

Figure 4. VLM-I scores across eight construction categories. Each cell reflects the average reasoning quality for a model-category pair.

spatial coordination. Claude-Sonnet-4.5 [3] and DeepSeek v3.1 [24] follow closely, especially in categories emphasizing geometric regularity and quantitative precision. Qwen3-VL [62] performs well in structurally constrained tasks like Basic Constructions and Triangle Properties but lags in proportion-based reasoning. On the other end, models such as GPT-4o [19], GLM-4.5V [17], and Qwen-Image [56] consistently underperform, with average scores below 23 across most categories, suggesting broad limitations in spatial grounding. Categories that demand accurate application of geometric theorems or strict measurement reasoning prove most difficult, as evidenced by widespread performance drops across the board. These results reinforce the need for construction-level benchmarking to surface fine-grained differences in multimodal reasoning competence.

4.4. Error Analysis

Figure 5 illustrates two typical examples that highlight reasoning and construction errors. Additional failure cases are provided in Appendix D.2. In the first example, the model misinterprets the constraint and produces the exact inverse of the intended geometry. Instead of constructing a rectangle inscribed *within* a circle, it generates a circle inscribed *inside* a square. In the second example, the textual instructions are logically correct and follow the ground-truth reasoning sequence. However, the corresponding visual outputs (Steps 2–4) are severely inconsistent with both the text and geometric principles. Key entities—such as the circle’s center O and the midpoint M of radius OA —are misplaced, and the perpendicular bisector is drawn incorrectly.

5. Conclusion

We introduce GGBench, a benchmark designed to evaluate geometric generative reasoning, addressing the gap between

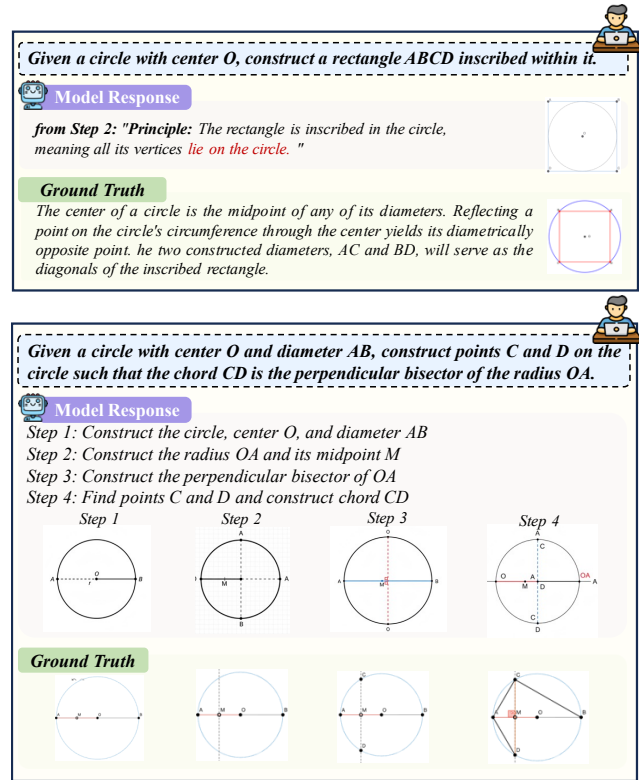


Figure 5. Error analysis of representative failure cases.

multimodal understanding and structured, verifiable generation. Through extensive experiments across both end-to-end UMMs and reasoning-grounded LLMs/LRMs, we observed a systematic gap between unconstrained visual generation and code-based constructive reasoning. GGBench establishes a rigorous and interpretable testbed for UMMs.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv e-prints*, pages arXiv-2303, 2023. 7, 19, 23, 25
- [2] 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. 1
- [3] Anthropic. Claude sonnet 4.5 system card. Technical report, Anthropic PBC, 2025. Official system card describing Claude Sonnet 4.5 capabilities and safety evaluation. Available at: <https://assets.anthropic.com/m/12f214efcc2f457a/original/Claude-Sonnet-4-5-System-Card.pdf>. 7, 8, 19, 22, 23, 24, 25
- [4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 3
- [5] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523, 2021. 3
- [6] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Januspro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 3
- [7] Konstantin Chernyshev, Vitaliy Polshkov, Ekaterina Artemova, Alex Myasnikov, Vlad Stepanov, Alexei Miasnikov, and Sergei Tilga. U-math: A university-level benchmark for evaluating mathematical skills in llms. *arXiv preprint arXiv:2412.03205*, 2024. 4
- [8] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 2
- [9] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blstein, 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. 3, 4, 7, 19, 23, 25
- [10] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 3, 7, 19, 24
- [11] Alisa Fortin, Guillaume Vernade, Kat Kampf, and Ammaar Reshi. Introducing Gemini 2.5 Flash Image: our state-of-the-art image generation and editing model. Google Developers Blog, 2025. 1, 3, 19
- [12] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjuan Zhong, Yufei Wang, Lanqing HONG, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. In *The Thirteenth International Conference on Learning Representations*, 2025. 1
- [13] Himanshu Gupta, Shreyas Verma, Ujjwala Ananthaswaran, Kevin Scaria, Mihir Parmar, Swaroop Mishra, and Chitta Baral. Polymath: A challenging multi-modal mathematical reasoning benchmark. *arXiv preprint arXiv:2410.14702*, 2024. 3, 7
- [14] Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, et al. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. *arXiv preprint arXiv:2504.11456*, 2025. 3
- [15] Zhitao He, Zongwei Lyu, Dazhong Chen, Dadi Guo, and Yi R Fung. Matp-bench: Can mllm be a good automated theorem prover for multimodal problems? *arXiv preprint arXiv:2506.06034*, 2025. 3
- [16] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021. 2
- [17] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Li-hang Pan, et al. Glm-4.5v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pages arXiv-2507, 2025. 7, 8, 19, 22, 23
- [18] Muye Huang, Lingling Zhang, Jie Ma, Han Lai, Fangzhi Xu, Yifei Li, Wenjun Wu, Yaqiang Wu, and Jun Liu. Charts-ketcher: Reasoning with multimodal feedback and reflection for chart understanding. *arXiv preprint arXiv:2505.19076*, 2025. 3
- [19] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1, 3, 7, 8, 19, 22, 23
- [20] Jiulin Li, Ping Huang, Yexin Li, Shuo Chen, Juewen Hu, and Ye Tian. A unified multi-agent framework for universal multimodal understanding and generation. *arXiv preprint arXiv:2508.10494*, 2025. 3
- [21] Xiaoyuan Li, Moxin Li, Wenjie Wang, Rui Men, Yichang Zhang, Fuli Feng, Dayiheng Liu, and Junyang Lin. Mathopeval: A fine-grained evaluation benchmark for visual operations of mllms in mathematical reasoning. *arXiv preprint arXiv:2507.18140*, 2025. 3
- [22] Yi Li, Haonan Wang, Qixiang Zhang, Boyu Xiao, Chenchang Hu, Hualiang Wang, and Xiaomeng Li. Unieval: Unified holistic evaluation for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.10483*, 2025. 3
- [23] Zhongzhi Li, Ming-Liang Zhang, Pei-Jie Wang, Jian Xu, Rui-Song Zhang, Yin Fei, Zhi-Long Ji, Jin-Feng Bai, Zhen-Ru Pan, Jiaxin Zhang, et al. Cmmath: A chinese multi-modal

- math skill evaluation benchmark for foundation models. In *ACL*, pages 2690–2726, 2025. 4
- [24] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *CoRR*, 2024. 7, 8, 19, 22, 23, 24, 25
- [25] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 2
- [26] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26439–26455, 2024. 1
- [27] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Øyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 2
- [28] 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. 2, 3, 4, 6, 7
- [29] Yanzuo Lu, Xin Xia, Manlin Zhang, Huafeng Kuang, Jianbin Zheng, Yuxi Ren, and Xuefeng Xiao. Hyper-bagel: A unified acceleration framework for multimodal understanding and generation. *arXiv preprint arXiv:2509.18824*, 2025. 3
- [30] Jingkun Ma, Runzhe Zhan, Derek F Wong, Yang Li, Di Sun, Hou Pong Chan, and Lidia S Chao. Visaidmath: Benchmarking visual-aided mathematical reasoning. *arXiv preprint arXiv:2410.22995*, 2024. 3
- [31] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7739–7751, 2025. 3
- [32] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruvi Shah, Xianzhi Du, Futang Peng, Anton Belyi, et al. Mm1: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pages 304–323. Springer, 2024. 3
- [33] Yuting Ning, Zhenya Huang, Xin Lin, Enhong Chen, Shiwei Tong, Zheng Gong, and Shijin Wang. Towards a holistic understanding of mathematical questions with contrastive pre-training. In *AAAI*, pages 13409–13418, 2023. 4
- [34] OpenAI. Gpt-5 system card. Technical report, OpenAI, 2025. Official system card document for GPT-5; available at: <https://cdn.openai.com/pdf/8124a3ce-ab78-4f06-96eb-49ea29ffb52f/gpt5-system-card-aug7.pdf>. 4, 7, 13, 19, 22, 23, 24, 25
- [35] Nilay Pande, Sahiti Yerramilli, Jayant Sravan Tamarapalli, and Rynaa Grover. Marvl-qa: A benchmark for mathematical reasoning over visual landscapes. *arXiv preprint arXiv:2508.17180*, 2025. 3
- [36] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024. 3, 4, 7
- [37] Hanoona Rasheed, Abdelrahman Shaker, Anqi Tang, Muhammad Maaz, Ming-Hsuan Yang, Salman Khan, and Fahad Shahbaz Khan. Videomathqa: Benchmarking mathematical reasoning via multimodal understanding in videos. *arXiv preprint arXiv:2506.05349*, 2025. 3
- [38] Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv e-prints*, pages arXiv–2509, 2025. 7, 19, 24, 25
- [39] Chengyu Shen, Zhen Hao Wong, Runming He, Hao Liang, Meiyi Qiang, Zimo Meng, Zhengyang Zhao, Bohan Zeng, Zhengzhou Zhu, Bin Cui, et al. Let’s verify math questions step by step. *arXiv preprint arXiv:2505.13903*, 2025. 3
- [40] Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See Kiong Ng, Lidong Bing, and Roy Lee. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. In *EMNLP*, pages 4663–4680, 2024. 4
- [41] Weikang Shi, Aldrich Yu, Rongyao Fang, Houxing Ren, Ke Wang, Aojun Zhou, Changyao Tian, Xinyu Fu, Yuxuan Hu, Zimu Lu, et al. Mathcanvas: Intrinsic visual chain-of-thought for multimodal mathematical reasoning. *arXiv preprint arXiv:2510.14958*, 2025. 3
- [42] Yuxin Song, Wenkai Dong, Shizun Wang, Qi Zhang, Song Xue, Tao Yuan, Hu Yang, Haocheng Feng, Hang Zhou, Xinyan Xiao, et al. Query-kontext: An unified multimodal model for image generation and editing. *arXiv preprint arXiv:2509.26641*, 2025. 3
- [43] Kai Sun, Yushi Bai, Ji Qi, Lei Hou, and Juanzi Li. Mm-math: Advancing multimodal math evaluation with process evaluation and fine-grained classification. *arXiv preprint arXiv:2404.05091*, 2024. 3, 6, 7
- [44] Cheng Tan, Jingxuan Wei, Zhangyang Gao, Linzhuang Sun, Siyuan Li, Ruifeng Guo, Bihui Yu, and Stan Z Li. Boosting the power of small multimodal reasoning models to match larger models with self-consistency training. In *European Conference on Computer Vision*, pages 305–322. Springer, 2024. 2
- [45] Trieu Trinh and Thang Luong. Alphageometry: An olympiad-level ai system for geometry. *Google DeepMind*, 17, 2024. 1
- [46] Junling Wang, Anna Rutkiewicz, April Yi Wang, and Mrinmaya Sachan. Generating pedagogically meaningful visuals for math word problems: A new benchmark and analysis of text-to-image models. *arXiv preprint arXiv:2506.03735*, 2025. 3, 7

- [47] 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. *NeurIPS*, 37:95095–95169, 2024. 3, 4
- [48] Ke Wang, Junting Pan, Linda Wei, Aojun Zhou, Weikang Shi, Zimu Lu, Han Xiao, Yunqiao Yang, Houxing Ren, Mingjie Zhan, et al. Mathcoder-vl: Bridging vision and code for enhanced multimodal mathematical reasoning. *arXiv preprint arXiv:2505.10557*, 2025. 3
- [49] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3
- [50] Peijie Wang, Zhong-Zhi Li, Fei Yin, Dekang Ran, and Cheng-Lin Liu. Mv-math: Evaluating multimodal math reasoning in multi-visual contexts. In *CVPR*, pages 19541–19551, 2025. 4
- [51] 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. 3, 7
- [52] Xiaokun Wang, Peiyu Wang, Jiangbo Pei, Wei Shen, Yi Peng, Yunzhuo Hao, Weijie Qiu, Ai Jian, Tianyidan Xie, Xuchen Song, et al. Skywork-vl reward: An effective reward model for multimodal understanding and reasoning. *arXiv preprint arXiv:2505.07263*, 2025. 3
- [53] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2
- [54] Jingxuan Wei, Caijun Jia, Qi Chen, Honghao He, Linzhuang Sun, Conghui He, Lijun Wu, Bihui Yu, and Cheng Tan. Geoint-r1: Formalizing multimodal geometric reasoning with dynamic auxiliary constructions. *arXiv preprint arXiv:2508.03173*, 2025. 1, 3
- [55] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *CVPR*, pages 12966–12977, 2025. 3, 7, 19, 24, 25
- [56] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv e-prints*, pages arXiv–2508, 2025. 7, 8, 19
- [57] Teng Xiao, Zuchao Li, and Lefei Zhang. Omnibridge: Unified multimodal understanding, generation, and retrieval via latent space alignment. *arXiv preprint arXiv:2509.19018*, 2025. 3
- [58] Yicheng Xiao, Lin Song, Rui Yang, Cheng Cheng, Zunnan Xu, Zhaoyang Zhang, Yixiao Ge, Xiu Li, and Ying Shan. Haploomni: Unified single transformer for multimodal video understanding and generation. *arXiv preprint arXiv:2506.02975*, 2025. 3
- [59] Wulin Xie, Yi-Fan Zhang, Chaoyou Fu, Yang Shi, Bingyan Nie, Hongkai Chen, Zhang Zhang, Liang Wang, and Tieniu Tan. Mme-unify: A comprehensive benchmark for unified multimodal understanding and generation models. *arXiv preprint arXiv:2504.03641*, 2025. 1
- [60] Wulin Xie, Yi-Fan Zhang, Chaoyou Fu, Yang Shi, Bingyan Nie, Hongkai Chen, Zhang Zhang, Liang Wang, and Tieniu Tan. Mme-unify: A comprehensive benchmark for unified multimodal understanding and generation models. *arXiv preprint arXiv:2504.03641*, 2025. 3
- [61] Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025. 3
- [62] 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 e-prints*, pages arXiv–2505, 2025. 7, 8, 19, 22, 23, 25
- [63] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 3
- [64] Xinwu Ye, Chengfan Li, Siming Chen, Wei Wei, and Xiangu Tang. Mmscibench: Benchmarking language models on chinese multimodal scientific problems. *arXiv preprint arXiv:2503.01891*, 2025. 3, 7
- [65] Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, et al. Internlm-math: Open math large language models toward verifiable reasoning. *arXiv preprint arXiv:2402.06332*, 2024. 3
- [66] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 2
- [67] 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 *CVPR*, pages 9556–9567, 2024. 3
- [68] Lingfei Zeng, Fengdi Che, Xuhan Huang, Fei Ye, Xu Xu, Binhang Yuan, and Jie Fu. Veriequivbench: An equivalence score for ground-truth-free evaluation of formally verifiable code. *arXiv preprint arXiv:2510.06296*, 2025. 3
- [69] Boning Zhang, Chengxi Li, and Kai Fan. Mario eval: Evaluate your math llm with your math llm—a mathematical dataset evaluation toolkit. *arXiv preprint arXiv:2404.13925*, 2024. 4
- [70] Jiaxin Zhang, Zhong-Zhi Li, Ming-Liang Zhang, Fei Yin, Cheng-Lin Liu, and Yashar Moshfeghi. Geoeval: Benchmark for evaluating llms and multi-modal models on geometry problem-solving. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1258–1276, 2024. 3, 7
- [71] 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 *ECCV*, pages 169–186. Springer, 2024. [3](#), [6](#), [7](#)
- [72] Xinjie Zhang, Jintao Guo, Shanshan Zhao, Minghao Fu, Lunhao Duan, Jiakui Hu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, et al. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*, 2025. [1](#), [3](#)
- [73] Yunhang Shen Yulei Qin Mengdan Zhang, Xu Lin Jinrui Yang Xiawu Zheng, Ke Li Xing Sun Yunsheng Wu, Ronrong Ji Chaoyou Fu, and Peixian Chen. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 18, 2021. [2](#)
- [74] Tianshi Zheng, Kelvin Kiu-Wai Tam, Newt Hue-Nam K Nguyen, Baixuan Xu, Zhaowei Wang, Jiayang Cheng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Tianqing Fang, et al. Newtonbench: Benchmarking generalizable scientific law discovery in llm agents. *arXiv preprint arXiv:2510.07172*, 2025. [3](#)
- [75] Minxuan Zhou, Hao Liang, Tianpeng Li, Zhiyu Wu, Mingan Lin, Linzhuang Sun, Yaqi Zhou, Yan Zhang, Xiaoqin Huang, Yicong Chen, et al. Mathscape: Evaluating mllms in multimodal math scenarios through a hierarchical benchmark. *arXiv preprint arXiv:2408.07543*, 2024. [3](#), [4](#), [6](#), [7](#)
- [76] Xianwei Zhuang, Yuxin Xie, Yufan Deng, Liming Liang, Jinghan Ru, Yuguo Yin, and Yuexian Zou. Vargpt: Unified understanding and generation in a visual autoregressive multimodal large language model. *arXiv preprint arXiv:2501.12327*, 2025. [3](#)
- [77] 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. [3](#)