MathLABS: Evaluation of Visual Math Reasoning Capabilities in Multimodal Large Language Model

Akhilesh Vangala ‡ **Bruce Zhang ‡** **Lucas Yao ‡** **Sahil Parupudi ‡**
Center for Data Science, New York University
New York, NY 10011 USA
{sv3129,yz8063,ly2808,sp9019}@nyu.edu

Abstract

The capabilities of Multimodal Large Language Model (MLLMs) have expanded rapidly in recent years. Their applications now span a wide spectrum, ranging from code generation with tools such as CoPilot and Cursor to video synthesis with systems such as SORA and Veo, making MLLMs increasingly ubiquitous. These models have also achieved record-breaking performance on rigorous benchmarks, despite these advances, a critical challenge in assessing their ability to perform sophisticated mathematical reasoning remains, particularly in settings where visual input plays a central role (P. Wang et al., 2025). This difficulty is especially pronounced in complex tasks that require the interpretation and manipulation of visual information, such as graphs, diagrams, and geometric figures, to derive precise mathematical solutions (Li et al., 2025). To this end, we propose to construct a compact, purpose-built visual–math dataset by sampling existing large-scale corpora into a small subset that preserves domain coverage, difficulty gradients, and known failure modes of models. We will also author a targeted set of synthetic questions, followed by light expert checks for validity and clarity. The resulting benchmark is purpose-built for relative evaluation, rigorous stress testing, and an efficiency study of dataset-driven workflows.

Introduction

Multimodal large language models (MLLMs) have recently achieved strong performance on a variety of visually conditioned mathematical benchmarks, yet they still exhibit notable failures on tasks that require fine-grained diagram understanding and multi-step visual reasoning. (Lu et al., 2021, 2024; K. Wang et al., 2024; Wu et al., 2024; Zhang et al., 2024). At the same time, a modern incarnation of Moravec's paradox is emerging: models that excel at complex, rule-based tasks such as advanced games can still struggle with intuitive, visually grounded skills that humans find effortless, and narrow puzzle-style benchmarks can give an overly optimistic picture of progress. (Karpathy, 2024)

A growing body of work has introduced benchmarks targeting visual mathematics, including diagram-centric evaluations and multi-image reasoning tasks, which collectively reveal systematic weaknesses in current MLLMs. FrontierMath further highlights that even state-of-the-art models remain far from expert-level performance on challenging mathematical problems, with reported accuracies well below human levels.(Glazer et al., 2024).More recent efforts such as MathFlow and VisioMath decompose visual mathematical reasoning into distinct perception and inference stages and design visually similar options for multiple-choice problems, offering a more controlled view of visual understanding but still leaving important gaps(.Chen et al., 2025; Hu et al., 2025)

In particular, existing benchmarks often underrepresent authentic discrete mathematics tasks, rely heavily on toy puzzles, lack systematic difficulty calibration, and provide limited structured metadata for rigorous analysis.(Lu et al., 2024; K. Wang et al., 2024; Wu et al., 2024) These limitations are at odds with the needs of small-data regimes, where careful curation, topic coverage, and difficulty control are critical for meaningful evaluation. To address these gaps, this work introduces a small-data benchmark of discrete mathematics problems with associated figures, combining textbook-derived and synthetically generated items under a unified JSON schema enriched with topic labels, difficulty levels, and cognitive-skill annotations, and uses this benchmark to systematically evaluate the visual reasoning capabilities of contemporary MLLMs.(Chen et al., 2025; Hu et al., 2025; Lee et al., 2025)

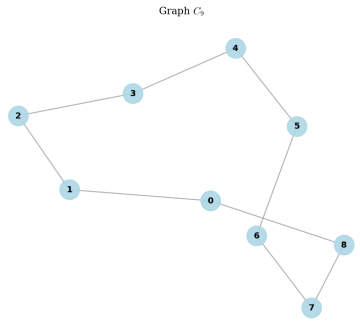

Figure 1: *Is the above graph a bipartite graph?*

The above figure shows a question from the MathLABS dataset

Related Work

Table 1 summarizes the major benchmarks and systems that define this landscape and highlights the core difficulties they expose for current MLLMs.

Table 1: Summary of major visual math benchmarks and systems.

Name (Year)	Key Contributions	Challenges Highlighted
MathVerse (2024) (Zhang et al., 2024)	Introduces vision-dependent math problems to test whether MLLMs genuinely use diagrams for reasoning.	Models struggle with fine-grained visual distinctions and consistent diagram interpretation.
MATH-Vision (2024) (K. Wang et al., 2024)	Large-scale benchmark with competition-style problems across 16 disciplines and explicit difficulty levels.	MLLMs often fail multi-step reasoning and harder competition-level tasks.
MathVista (2024) (Lu et al., 2024)	Evaluates math reasoning in rich visual contexts, combining diverse problem types and modalities.	Context-heavy visuals create ambiguity; models show inconsistent grounding.
VCBench (2024) (Wu et al., 2024)	Multi-image visual reasoning benchmark with explicit visual dependencies.	Difficulties in aggregating information across images and resolving spatial relations.
Geometry3K (2021) (Lu et al., 2021)	Geometry benchmark with annotated diagrams and interpretable solution structures.	Models struggle with geometric constraints and precise diagram-to-text mapping.
FrontierMath (2024) (Glazer et al., 2024)	Collection of expert-curated, research-level math problems showing gaps relative to human reasoning.	Even strong MLLMs achieve low accuracy, exposing deep reasoning limitations.
MathFlow (2025) (Chen et al., 2025)	Modular perception–inference pipeline that converts diagrams into enriched textual descriptions.	Perception errors propagate; fine-grained visual extraction remains unreliable.
VisioMath (2025) (Hu et al., 2025)	Multiple-choice benchmark with visually similar diagrams to test fine-grained comparison.	Models exhibit positional bias and failures in subtle diagram comparison.

Methodology

The proposed benchmark is constructed around figure-based discrete mathematics questions, where each item either includes an essential diagram component or is fully rendered as an image prompt for the model.Problems are primarily drawn from standard discrete mathematics textbooks and are complemented with synthetically generated questions to broaden topic and difficulty coverage. All questions are normalized into a unified JSON schema that includes the problem statement, figure metadata, correct answer, and annotations for topic, difficulty, and cognitive skill level.

Question Construction Pipeline

Figures, concepts, and base questions are first extracted from reference materials such as "Discrete Mathematics and Its Applications," and then programmatically extended or modified to satisfy the schema and rubric.(Rosen, 2012) Additional graphs and diagrams are generated using tools such as `matplotlib` and `NetworkX`, enabling systematic variation over graph size, structure, and visual layout while maintaining alignment with discrete math concepts like graph connectivity, combinatorics, and probability.(Lee et al., 2025) The resulting pool of problems is split into extracted and generated subsets, which are subsequently sampled to assemble evaluation sets with balanced topic and difficulty distributions.

To match the presentation pipeline (Database $\rightarrow$ Questions $\rightarrow$ Models), the benchmark construction and evaluation process is summarized in.A centralized database stores all question instances, metadata, and path to the associated figures; evaluation batches are then sampled from this database and dispatched to multiple models under a standardized interface.

Model Evaluation and Validation

For evaluation, a set of contemporary MLLMs is selected, covering both proprietary and open-source families with vision capabilities. In each run, the system samples a batch of questions, renders the corresponding figures, and prompts each model with a standardized template that includes the image and a concise textual instruction, recording the model's answer and response time. This follows a pre-defined rubric.A master(Gemini 2.5 Flash)(Comanici et al., 2025) model is used as an automated validator to cross-check answers and flag potential annotation issues, while the student" models constitute the main evaluation targets. All results are stored in the database and analyzed through a dashboard that supports breakdowns by difficulty, topic, question source (extracted vs. generated), model family, and run index.

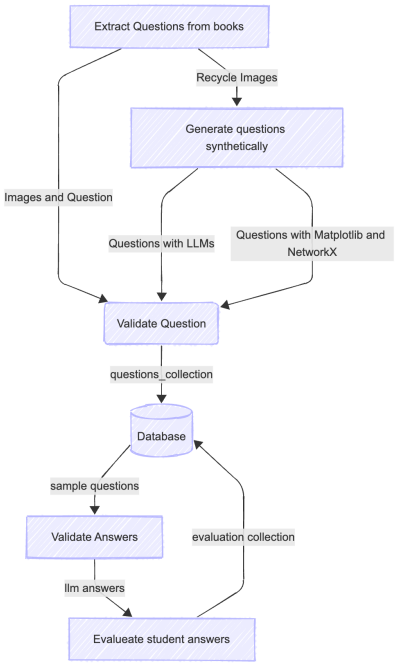

Figure 2: *Process Flow*

Experiments

Dataset

All experiments in the study were conducted on the finalized MathLabs question set, which consists of multiple-choice mathematics problems drawn from 13 major mathematical subjects, including number theory, discrete mathematics, graph theory, probability, calculus, mathematical logic, geometry, and computer science. Each problem includes a fixed set of 3 distractors and an accompanying diagram when applicable. The distractors are designed to mimic realistic mathematical errors, including common computational slips, conceptual misunderstandings, procedural mistakes, or visual or spatial misinterpretations to the diagrams. These distractors are intentionally plausible, ensuring that accuracy reflects genuine reasoning ability rather than trivial elimination.

The full dataset combines extracted questions from authoritative sources (180, 39.1%) and validated model-generated items (280, 60.9%). For this experiment, we sampled from the full question set ($N = 460$).

Models Evaluated

We evaluated a representative set of contemporary multimodal LLMs covering a wide spectrum of model sizes and families. The models included frontier proprietary systems (e.g., grok-4.1-fast, ERNIE-4.5-PT), mid-sized open-source vision-language models (e.g., gemma-3-27b-it, GLM-4.1V-Thinking, nemotron-nano-vl), and smaller instruction-tuned baselines (e.g., mistral-small-instruct, qwen2.5-vl-7b). All models were queried through their official API endpoints using default inference settings.

Evaluation Tasks and Metrics

We evaluate models on four dimensions:

- Accuracy: percentage of correctly predicted answers.
- Difficulty-conditioned accuracy: performance split across easy, medium and ?HARD? questions from the dataset.
- Topic-wise performance: accuracy and average response time grouped by mathematical topic.
- Response latency: end-to-end model inference time per question, measured at the API level.

For accuracy metrics, 95% accuracy confidence intervals were obtained using the standard normal approximation to the binomial proportion estimator. The Topic-level analyzes were restricted to topics with at least 3 questions.

Implementation Details

All evaluations were executed through our batch runner with a batch size of 10. Before each request, the order of answer choices was randomly shuffled. The system logged the model's selected answer, free-form reasoning, latency, predicted difficulty label, and metadata for downstream aggregation. All computations and visualizations in the Results section were organized using our Streamlit dashboard, which aggregates prediction traces stored in our database.

Results

Overall Model Performance

Figure 3 and Table 2 summarize the accuracy and 95% confidence intervals across all evaluated models. Among these models, grok-4.1-fast demonstrates the strongest overall performance, clearly outperforming both proprietary and open-source baselines on the benchmark. A second tier of models, including gemma-3-27b-it and ERNIE-4.5-PT exhibits competitive accuracy levels, though with notably wider confidence intervals, indicating greater variability in their responses. Mid-sized vision-language models such as Mistral-small-instruct and sherlock-dash-alpha form a middle cluster with moderate performance. In contrast, smaller instruction-tuned models (e.g., qwen-2.5-vl variants) consistently underperform, reflecting more limited visual-mathematical reasoning capacity.

Difficulty-Based Performance

Table 3 shows that accuracy decreases with increasing problem difficulty. Models perform well on EASY items with accuracies concentrated in higher ranges, while MEDIUM questions introduce lower mean accuracy with greater dispersion. HARD questions produce near-floor performance. This gradient demonstrates that the models are still incapable to robustly address the increasing difficulty and visual reasoning demands for the questions.

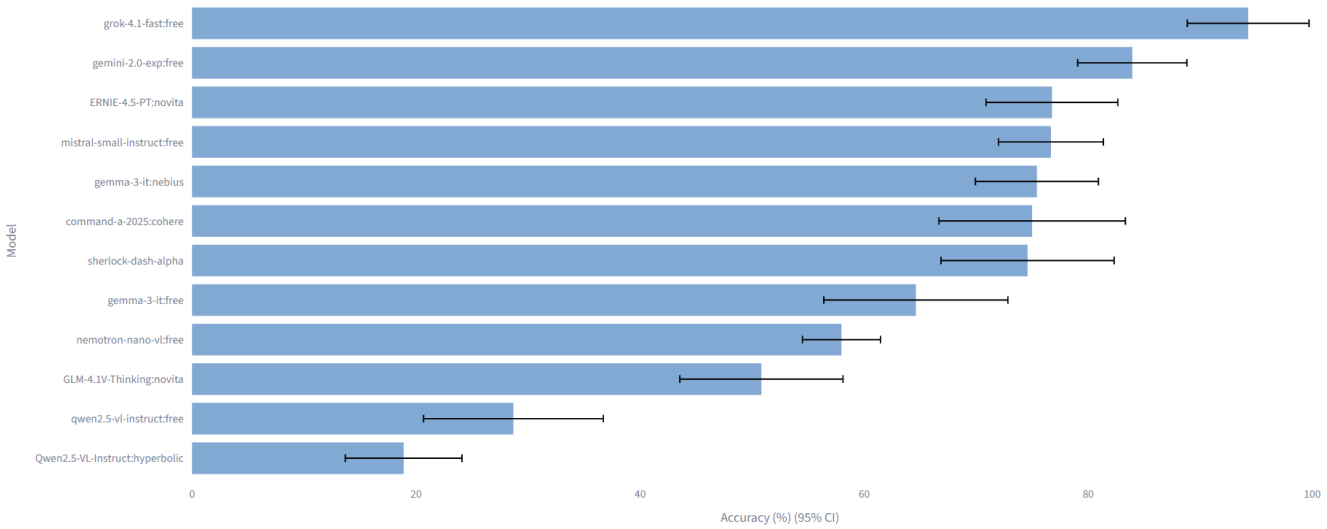

Figure 3: *Model Performance Comparison (with 95% Confidence Intervals*

Table 2: Overall model accuracy and latency with 95% confidence intervals.

Model	**Accuracy (%)**	**Latency (seconds)**
grok-4.1-fast-free	94.29 ± 5.44	14.72 ± 4.87
gemini-2.0-exp-free	83.94 ± 4.87	3.12 ± 1.07
ERNIE-4.5-PT-novita	76.77 ± 5.88	5.29 ± 0.57
mistral-small-instruct	76.68 ± 4.68	3.19 ± 0.32
gemma-3-27b-it-nebius	75.42 ± 5.49	1.64 ± 0.18
command-a-2025-cohere	75.00 ± 8.32	7.16 ± 4.49
sherlock-dash-alpha	74.59 ± 7.73	1.68 ± 0.18
gemma-3-it-free	64.62 ± 8.22	1.64 ± 0.18
GLM-4.1V-Thinking-novita	50.83 ± 7.28	9.50 ± 1.81
Qwen2.5-VL-instruct	28.69 ± 8.03	4.11 ± 1.14
Qwen2.5-VL-Instruct-hyperbolic	18.89 ± 5.21	3.99 ± 0.36

Table 3: Mean Accuracies Based on Difficulty Levels

Model	**Mean Accuracy (%)**
EASY	66.13
MEDIUM	49.20
HARD	12.50

Response Time Comparison

Table 2 compares the average latency with their 95% confidence intervals. The results reveal a clear hierarchy and architectural complexity. grok-4.1-fast demonstrates the longest response times reflect the heavier reasoning operations in comparison to other models. Mid-sized models like GLM-4-Thinking form a second tier with moderate latency. The latency pattern highlights the trade-off between reasoning capability and computational cost, that models with stronger performance on complex tasks tend to have higher inference latency.

Topic-Level Performance

Figure 4 presents topic-level accuracy and average response time across all evaluated mathematical domains. Accuracy varies substantially by topic: computer science, topology, and probability achieve the highest correctness rates, followed by calculus and logic. In contrast, coordinate vectors, graph theory, and linear algebra exhibit notably lower accuracy, indicating that models struggle with tasks requiring geometric or structural reasoning.

Inference time shows a different but equally topic-dependent pattern. Geometry and coordinate vectors yield the longest average latencies, often exceeding 9–10 seconds,suggesting increased computational difficulty or more complex visual parsing in these domains. Topics such as mathematical logic, calculus, and topology produce the shortest response times.

Together, these results indicate that model performance is

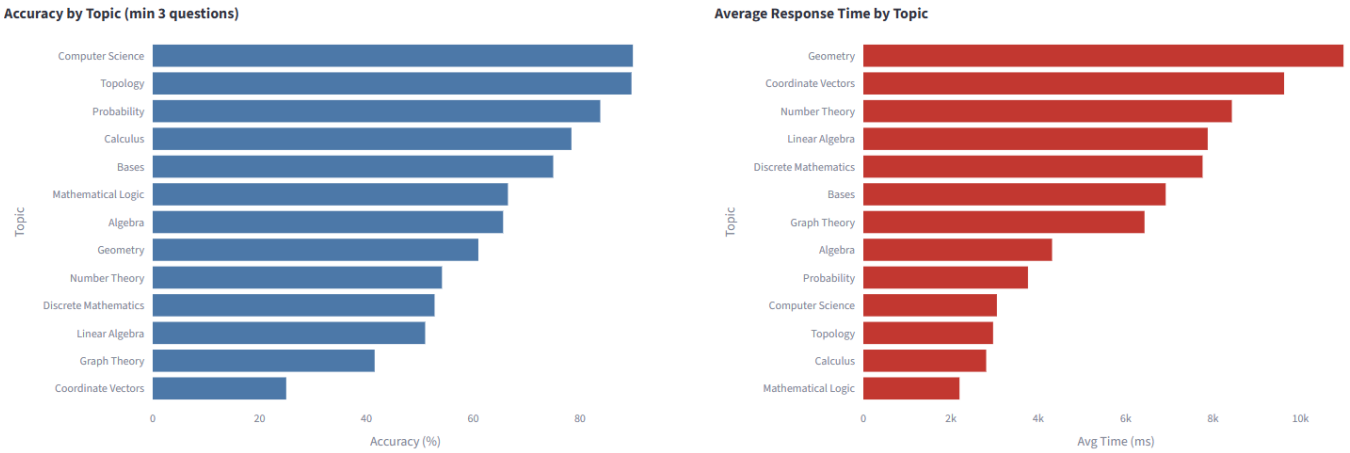

Figure 4: *Topic Performance Comparison*

not uniform across mathematical areas: models excel in symbolic and discrete reasoning domains, while geometric, spatial, and vector-based problems remain the most challenging both in accuracy and inference efficiency.

Conclusion

We introduced a small-data benchmark designed to probe the visual reasoning abilities of multimodal large language models (MLLMs) on discrete mathematics problems with explicit figure components. By combining textbook-derived and synthetically generated items under a unified schema with rich metadata, the benchmark supports fine-grained analyses across topics, difficulty levels, and cognitive skills, while remaining feasible to construct and maintain in low-data regimes.

Empirical evaluation across a diverse set of MLLMs reveals substantial variability in performance by topic and difficulty. Models generally perform better on generated, easier questions but struggle on harder, authentic problems and those requiring multi-step reasoning over diagrams. These results align with broader trends in existing visual math benchmarks, highlighting that current MLLMs still fall short of robust, human-like visual mathematical reasoning, particularly in discrete domains demanding precise interpretation of graph- and combinatorics-based figures

Future Work

The insights obtained from this research point toward several critical and promising avenues for continued investigation, both in the extension of the MathLabs benchmark and in the broader analysis of visual mathematical reasoning systems.

Firstly, to enhance the stability and generalizability of our findings, future work should prioritize **increased evaluation robustness**. This entails conducting a larger number of evaluation runs, particularly for the most visually complex and "HARD" difficulty-level questions. This effort will yield more stable performance estimates and enable a more precise, fine-grained comparison across diverse model architectures and training paradigms. The dataset also has to be tested on more state-of-the-art foundation models such as Gemini 3 and even GPT 5.2

Secondly, a crucial next step is the **incorporation of human performance baselines** on a curated subset of the MathLabs questions. This will provide necessary context for the reported model accuracies, allowing researchers to distinguish between MLLM failures on problems humans find trivial and failures on items that represent genuinely difficult mathematical challenges.

Thirdly, significant value lies in **extending the benchmark to include interactive modalities**. Future evaluations should explore how models perform when provided with mechanisms such as step-by-step solutions, scratchpad reasoning, or targeted, interactive hints. This approach will be instrumental in illuminating the cognitive gap between current MLLM reasoning and the iterative, conceptual problem-solving demonstrated by humans in discrete mathematics.

Finally, a promising direction involves shifting focus from consumption to **generative visual task evaluation**. Future research should aim to evaluate and enhance MLLMs' ability not only to interpret but also to *generate* mathematically correct and pedagogically useful diagrams from symbolic or natural language representations. This would leverage recent advances in vector-based diagram generation and push the boundaries of MLLMs toward becoming truly versatile visual mathematical assistants.

References

Chen, F., et al. (2025). Mathflow: Enhancing the perceptual flow of mllms for visual mathematical problems. *arXiv preprint*.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., Marris, L., Petulla, S., Gaffney, C., Aharoni,

A., Lintz, N., Pais, T. C., Jacobsson, H., Szpektor, I., Jiang, N.-J., … Helmholz, W. (2025). Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

Glazer, E., et al. (2024). Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint*.

Hu, X., et al. (2025). Visiomath: Benchmarking figure-based mathematical reasoning in large multimodal models. *arXiv preprint*.

Karpathy, A. (2024). On what is hard for humans vs. machines [Accessed: 2025-12-02].

Lee, J., et al. (2025). From text to visuals: Using llms to generate math diagrams with vector graphics. *arXiv preprint*.

Li, C., Zhang, T., Wang, M., & Huang, H. (2025). VisioMath: Benchmarking Figure-based Mathematical Reasoning in LMMs. *arXiv e-prints*, Article arXiv:2506.06727, arXiv:2506.06727.

Lu, P., et al. (2021). Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *Proceedings of a major AI conference*.

Lu, P., et al. (2024). Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint*.

Rosen, K. H. (2012). *Discrete mathematics and its applications* (7th). McGraw-Hill.

Wang, K., et al. (2024). Math-vision: Measuring multimodal mathematical reasoning. *arXiv preprint*.

Wang, P., Li, Z.-Z., Yin, F., Yang, X., Ran, D., & Liu, C.-L. (2025). MV-MATH: Evaluating Multimodal Math Reasoning in Multi-Visual Contexts. *arXiv e-prints*, Article arXiv:2502.20808, arXiv:2502.20808.

Wu, J., et al. (2024). Evaluating multimodal math reasoning in multi-visual contexts. *arXiv preprint*.

Zhang, R., et al. (2024). Mathverse: Does your multi-modal llm truly see the diagrams? *arXiv preprint*.

Source Code : https://github.com/mathlabsNYU/mathlabs

Image Dataset : https://huggingface.co/datasets/brucezhang41/MathLABS

The authors acknowledge the usage of Perplexity, an AI search engine model developed by Perplexity AI, in the preparation of this assignment. Perplexity was employed in the following manners within this assignment: Brainstorming, Code Debugging, Grammar Correction, Latex Formatting

AI assistance was also employed in the Methodology, in the System Design and Database Management Processes