Beyond Seeing: Alternative Representations for Image-Dependent Mathematics

Ethan Croteau
Worcester Polytechnic Institute
Worcester, MA, USA
ecroteau@wpi.edu

Neil Heffernan
Worcester Polytechnic Institute
Worcester, MA, USA
nth@wpi.edu

ABSTRACT

Multimodal large language models (MLLMs) are increasingly proposed for tutoring, feedback, and other educational uses, yet they remain brittle on mathematics problems that require interpreting diagrams, graphs, number lines, and other figures. This dissertation investigates whether many of these failures are fundamentally representational: models may have the mathematical knowledge needed to solve a problem, but fail because the visual information is not encoded in a sufficiently faithful form for downstream reasoning. To address this challenge, I study alternative representations of image-dependent mathematical content, including high-quality alt-text, structured textual descriptions, and executable Python reconstructions of the original figure. I first analyze difficult image-dependent middle-school mathematics problems to identify recurring visual bottlenecks in MLLM performance. I then evaluate whether alternative representations improve not only problem solving, but also educationally meaningful tasks such as hint generation, explanation, scaffolding, accessibility support, and related-problem creation. The dissertation contributes both an empirical account of representational failure in multimodal educational AI and a design framework for making visual mathematical content more usable in learning environments.

Keywords

multimodal large language models, mathematics education, visual representations, accessibility, educational AI

1. INTRODUCTION

Multimodal large language models (MLLMs) are increasingly being explored for tutoring, feedback, problem solving, and other educational applications. In mathematics education, however, many tasks depend on diagrams, graphs, geometric figures, tables, and other visuals that provide essential problem context. These are not simply text questions with accompanying illustrations; rather, the visual representation often contains information required to determine the correct answer or construct a valid solution. Such visually grounded tasks also arise in classroom-authentic tutoring and assessment settings, including platforms such as ASSISTments [3].

This challenge is especially important for educational data mining and AI in education because model success on visually grounded tasks depends on more than symbolic reasoning alone. A model may appear mathematically capable while still failing to count objects accurately, read a scale, bind labels to the correct visual elements, or preserve spatial relations needed for reasoning. As a result, strong performance on broad multimodal benchmarks can obscure weaknesses that matter in authentic learning environments [4]. Recent work on authentic middle-school image-dependent mathematics suggests that such failures are often driven by visual misreading rather than symbolic mathematics alone [2].

My dissertation is motivated by the hypothesis that many of these failures are fundamentally representational. In other words, the model may possess the mathematical reasoning required once the relevant information is available, but fail because the original image is not transformed into a sufficiently faithful internal or external representation for downstream reasoning. This dissertation therefore asks whether alternative representations of visual mathematical content can make image-dependent mathematics more usable for MLLMs in educational settings. Rather than treating the original image as the only viable input, I investigate forms such as high-quality alt-text, structured textual descriptions, and executable Python scripts that reconstruct the figure while preserving its mathematical structure.

2. PROBLEM AND RESEARCH GAP

MLLMs have become a central topic in educational technology research, with growing interest in their use for tutoring, automated feedback, accessibility, and content generation. At the same time, prior work in mathematics education and the learning sciences has shown that representations are not incidental to learning; they shape interpretation, inference, and strategy selection [1, 5]. This makes image-dependent mathematics an important testbed for educational AI.

Existing multimodal work often focuses on benchmark accuracy or broad demonstrations of capability [4]. While this work is valuable, it does not fully explain why visually grounded mathematics problems remain difficult in authentic educational settings, especially when success depends on

Ethan Croteau, and Neil Heffernan. Beyond Seeing: Alternative Representations for Image-Dependent Mathematics. In Anthony Botelho, Maria Mercedes T. Rodrigo, Adish Singla, Hiroaki Ogata, Hyojeong So, and Young Hoan Cho (eds.) Proceedings of the 19th International Conference on Educational Data Mining, Seoul, Republic of Korea, June, 2026, pp. 832–834. International Educational Data Mining Society (2026).

https://doi.org/10.5281/zenodo.21040024

precise visual interpretation. In such cases, failure may reflect not only reasoning limitations but also bottlenecks in how task-relevant visual information is represented.

A second gap is that most work treats the original image as the primary or only form of visual input. In educational systems, however, the same mathematical content can often be expressed in multiple forms. A diagram may be represented as alt-text, as a structured description of entities and relations, or as an executable program that reconstructs the figure. These alternatives may preserve different aspects of the original content and may vary in how well they support reasoning, explanation, accessibility, and content generation. Accessibility guidance emphasizes that such alternatives must preserve task-relevant structure, not merely summarize appearance [6, 7]. Despite this promise, representation choice has not been studied systematically for image-dependent mathematics in educational contexts.

3. RESEARCH QUESTIONS AND THEORETICAL FRAMING

The dissertation is organized around three questions:

RQ1: What kinds of image-dependent mathematics problems remain especially difficult for MLLMs, and what recurring visual bottlenecks contribute to these failures?

RQ2: To what extent can alternative representations of visual mathematical content improve MLLM understanding and problem solving on image-dependent tasks?

RQ3: Which forms of alternative representation are most useful for downstream educational tasks such as hint generation, explanation, scaffolding, accessibility support, and the creation of related practice problems?

These questions reflect a dissertation arc that moves from *diagnosis* to *intervention* to *educational application*. The central theoretical claim is that representation is not merely a delivery format for visual information; it is part of the mechanism by which a model can or cannot reason successfully about a task. Accordingly, this dissertation treats alternative representation as both an analytic lens for explaining multimodal failure and a design space for building more useful educational AI systems.

4. CURRENT PROGRESS

A substantial portion of the dissertation's diagnostic phase is already complete. Earlier work on authentic middle-school mathematics problems in which the image is required for solving showed that even strong contemporary MLLMs remain brittle when the figure contains task-critical information not recoverable from text alone. Across repeated evaluations with and without the figure, these studies suggested that visual misreading is a dominant failure mode, with recurring errors involving counting, measurement, label binding, and geometric relation extraction.

This prior result motivates the dissertation's shift from diagnosing image-dependent failure to studying representation as an intervention. Rather than viewing each failure only as evidence of insufficient model capability, I ask whether the same visual information can be expressed in forms that are more accessible to the model while remaining faithful to the underlying mathematics. To explore this possibility, I conducted early experiments using prompting to generate Python scripts that reconstruct selected images from difficult mathematics problems. These scripts, typically implemented with plotting or drawing libraries, serve as executable representations of the original figure. Unlike a static image, they can be inspected, edited, parameterized, and potentially reused for multiple educational purposes. Building on these experiments, I have also begun implementing an early-stage web prototype that converts structured math-task inputs into schema-constrained visual specifications, deterministic scene representations, executable Matplotlib code, and SVG outputs. The prototype logs intermediate artifacts and run-level metadata, enabling inspection and comparison of alternative representations, but it should currently be understood as research infrastructure rather than as a fully evaluated educational intervention.

These early experiments suggest that executable representations may have value beyond answer production. Because they expose the structure of a figure in a machine-readable and human-interpretable form, they may support hint generation, explanation, scaffolding, and the creation of related-but-not-identical problems. Likewise, alt-text and structured descriptions may improve accessibility for learners who cannot fully access the original image while also providing models with more explicit task-relevant information.

5. PROPOSED METHODOLOGY

The next phase of the dissertation will systematically compare multiple representations of the same image-dependent mathematical content. I currently anticipate studying at least four representation conditions: (1) the original image, (2) high-quality alt-text, (3) structured textual descriptions designed to preserve task-relevant visual relations, and (4) executable Python reconstructions of the original figure. Hybrid conditions that combine these forms may also be explored.

These representations will be evaluated across several educationally meaningful tasks. The first is direct problem solving on image-dependent mathematics items. Here I will measure whether alternative representations improve model performance and reduce specific types of visual failure. The second is hint and explanation generation, where the goal is not simply to produce a correct answer but to generate support that is instructionally useful and grounded in the relevant visual structure. The third is scaffolding and accessibility support, including whether alternative representations help make visual mathematical information more understandable to both models and learners. The fourth is related-problem generation, where executable or structured representations may enable the creation of new problems that are similar in concept but not identical in surface form.

Methodologically, I plan to combine quantitative evaluation with qualitative analysis. Quantitative measures will include problem-solving accuracy, consistency across repeated trials, and performance differences across representation conditions. Qualitative analyses will examine which visual bottlenecks persist, which are reduced, and what kinds of down-

stream outputs become more educationally useful under alternative representations. For generative tasks such as hints or related-problem creation, expert review may also be used to assess mathematical fidelity, pedagogical usefulness, and alignment with the intended concept.

6. EXPECTED CONTRIBUTIONS AND RELEVANCE TO EDM

This dissertation aims to contribute at three levels. First, it will provide an empirical characterization of the visual bottlenecks that make some image-dependent mathematics problems especially difficult for MLLMs. Second, it will contribute a methodological framework for studying alternative representations of visual mathematical content and their effects on downstream reasoning. Third, it will offer practical guidance for educational applications, including tutoring, accessibility support, grounded explanation, and content generation.

The work is relevant to EDM in several ways. It studies how representational choices shape model behavior on authentic educational tasks; it addresses human factors and explainability by examining why models fail on visually grounded items; and it connects to the broader EDM interest in comparing human and artificial intelligence on learning-relevant tasks. More broadly, the dissertation is intended to speak to both learning sciences and computer science perspectives by combining attention to representational meaning, learner accessibility, and instructional usefulness with rigorous evaluation of model performance.

7. FEEDBACK SOUGHT

I would especially value feedback in three areas. First, I seek guidance on how best to position the dissertation's central contribution: as a study of multimodal failure, a study of alternative representation, or a broader effort to support educational uses of visual mathematical content. Second, I seek feedback on the design of the representation comparisons, including which representation conditions and evaluation tasks should be prioritized for the dissertation. Third, I would value advice on how to strengthen the link between EDM, AIED, and learning sciences perspectives when evaluating educational usefulness beyond answer accuracy.

Overall, this dissertation aims to understand why "seeing" is often not enough for current MLLMs on image-dependent mathematics, and how alternative representations may help bridge that gap in ways that are meaningful for learners, educators, and educational AI systems.

8. ACKNOWLEDGMENTS

We would like to thank NSF (e.g., 2118725, 2118904, 1950683, 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, & 1535428), IES (e.g., R305N210049, R305D210031, R305A170137, R305A170243, R305A180401, & R305A120125), GAANN (e.g., P200A180088 & P200A150306), EIR (U411B190024 & S411B210024), ONR (N00014-18-1-2768), NIH (R44GM146483), and Schmidt Futures. None of the opinions expressed here are those of the funders.

9. REFERENCES

[1] S. Ainsworth. Deft: A conceptual framework for considering learning with multiple representations. *Learning and Instruction*, 16(3):183–198, 2006.

[2] E. Croteau and N. Heffernan. Seeing is solving: MLLMs, reasoning, and refusal in visual math. *Journal of Educational Data Mining*, 18(1):244–285, 2026.

[3] N. T. Heffernan and C. L. Heffernan. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4):470–497, 2014.

[4] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, Vienna, Austria, 2024. OpenReview.net.

[5] R. E. Mayer. *Multimedia Learning*. Cambridge University Press, Cambridge, UK, 3rd edition, 2020.

[6] W3C Accessibility Guidelines Working Group. Web content accessibility guidelines (wcag) 2.2. W3C Recommendation, Dec. 2024. Latest version: `https://www.w3.org/TR/WCAG22/`.

[7] W3C Web Accessibility Initiative. Images tutorial: Complex images. `https://www.w3.org/WAI/tutorials/images/complex/`, 2022. Updated 17 January 2022. Accessed 10 April 2026.