Explain with Visual Keypoints Like a Real Mentor! A Benchmark for Multimodal Solution Explanation

Jaewoo Park[1]*, Jungyang Park[1,2]*, Dongju Jang[1], Jiwan Chung[1], Byungwoo Yoo[2], Jaewoo Shin[2], Seonjoon Park[2], Taehyeong Kim[2], Youngjae Yu[3]

[1]Yonsei University
[2]Mathpresso
[3]Seoul National University
jerife@yonsei.ac.kr, youngjaeyu@snu.ac.kr

Abstract

With the rapid advancement of mathematical reasoning capabilities in Large Language Models (LLMs), AI systems are increasingly being adopted in educational settings to support students' comprehension of problem-solving processes. However, a critical component remains underexplored in current LLM-generated explanations: multimodal explanation. In real-world instructional contexts, human tutors routinely employ visual aids, such as diagrams, markings, and highlights, to enhance conceptual clarity. To bridge this gap, we introduce the *multimodal solution explanation* task, designed to evaluate whether models can identify visual keypoints, such as auxiliary lines, points, angles, and generate explanations that incorporate these key elements essential for understanding. To evaluate model performance on this task, we propose ME2, a multimodal benchmark consisting of 1,000 math problems annotated with visual keypoints and corresponding explanatory text that references those elements. Our empirical results show that current models struggle to identify visual keypoints. In the task of generating keypoint-based explanations, open-source models also face notable difficulties. This highlights a significant gap in current LLMs' ability to perform mathematical visual grounding, engage in visually grounded reasoning, and provide explanations in educational contexts. We expect that the multimodal solution explanation task and the ME2 dataset will catalyze further research on LLMs in education and promote their use as effective, explanation-oriented AI tutors.

Archive — https://me2-benchmark.github.io

1 Introduction

The traditional one-to-many educational model (i.e., one teacher for multiple students) is gradually transitioning to one-to-one personalized tutoring systems and online learning (Mukul and Büyüközkan 2023). Recent developments in Multimodal Large Language Models (MLLMs) have opened new opportunities for effective learning, such as estimating question difficulty (Park et al. 2024), assisting teachers in curriculum planning (Hu et al. 2024), and supporting interactive tutoring systems (Chevalier et al. 2024). In particular, numerous studies (Liu et al. 2023; Uesato et al. 2022; Lu et al. 2023) have focused on enhancing the mathematical reasoning abilities of MLLMs. As a result, MLLMs have led many students to use them as tools when faced with mathematical questions (Pardos and Bhandari 2024).

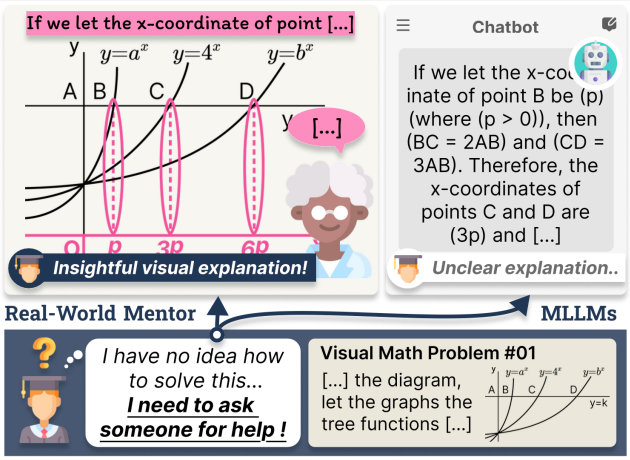

Figure 1: A student solving a math problem often benefits from visual cues—such as lines, symbols, or highlights—that human instructors use to aid understanding, unlike current AI models that focus solely on textual solutions. To serve as effective educational assistants, machines must go beyond answer generation and emulate human-like explanation strategies by explicitly incorporating and referencing visual elements.

However, from a student's perspective, relying solely on the reasoning footprints of MLLMs may not always be the best way to understand problems (Pardos and Bhandari 2024; Jia et al. 2024). One might wonder what distinguishes a broadly comprehensible explanation from a solution that merely yields the correct answer, for either a human or a model? A critical factor is the use of visual cues.

In actual educational settings, Dual Coding Theory (DCT) naturally occurs, providing effective learning opportuni-

*These authors contributed equally.

ties for students (Paivio 2013, 1990). According to DCT, combining verbal and visual information enhances student comprehension (Clark and Paivio 1991). As illustrated in Figure 1, human mentors often use visual scaffolding, such as annotated diagrams or highlighted keypoints on a blackboard, to foster intuitive understanding (Arcavi 2003; Stylianou 2010; Lee, Park, and Park 2024). In contrast, current AI models lack the capacity to generate such visual explanations. Moreover, existing datasets (Hendrycks et al. 2021; Lu et al. 2023; Wang et al. 2024a) focus solely on problem-solving and overlook educational objectives, making them insufficient for developing models capable of providing such forms of multimodal instructional support.

To address these limitations, we introduce *multimodal solution explanation*, a novel task that aims to enhance models' capacity to generate educationally effective and visually grounded mathematical explanations. In this task, models are required to (1) identify visual keypoints that are not present in the original problem but are crucial for understanding (e.g., lines, angles, annotations), and (2) generate explanatory text that explicitly refers to them. To benchmark performance on multimodal solution explanation, we propose Multimodal Explanations for Mathematics Education (**ME2**) benchmark. The ME2 includes not only the problem and its solution, but also annotations of the visual keypoints that serve as visual cues necessary to explain the solution, as well as keypoint-based explanatory texts aligned with them. Notably, we emphasize that ME2 goes beyond simple problem-solving, offering educational value in addressing the previously unexplored dimension of multimodal solution explanation.

Experiments on ME2 demonstrate that current MLLMs struggle to reliably identify visual keypoints. While closed-source models show potential in generating explanations grounded in visual keypoints, open-source generalist and math-specialized models show limited ability in this aspect. This suggests that current models largely fail to achieve robust mathematical visual grounding and visually grounded reasoning in educational contexts. We believe that ME2 will catalyze research toward strengthening mathematical visual grounding and reasoning, and advancing models that can serve as effective and student-friendly educational mentors.

Our contributions are as follows:

1. **A multimodal solution explanation task** that supports students' educational comprehension by identifying critical visual keypoints and generating explanatory text that explicitly references them.

2. **A ME2 benchmark**, rooted in authentic educational contexts, to rigorously assess multimodal solution explanation performance and facilitate further research on model-based explanations in real-world settings.

3. **Extensive experimental evaluations** of state-of-the-art MLLMs on multimodal solution explanation task, highlighting current limitations in recognizing and leveraging crucial visual keypoints to support effective learning.

2 Related Works

Language Models for Education. Recent advances in Large Language Models (LLMs) (Brown et al. 2020) have sparked significant interest in educational applications, particularly in personalized problem recommendation (Park et al. 2024), automated tutoring (Chevalier et al. 2024), and the provision of tailored feedback and customized curricula (Hu et al. 2024; Macina et al. 2023; Feng, Wang, and Sun 2023). For effective education, research suggests that combining textual and visual information enhances comprehension and memory more effectively than using text alone (Arcavi 2003; Stylianou 2010; Lee, Park, and Park 2024). As Clark and Paivio (1991) explains, dual representations supply multiple retrieval cues and cultivate richer mental models. Building on this insight, we introduce the multimodal solution explanation task to enable LLMs to offer learners more comprehensive learning opportunities. This task enables the model to pinpoint the visual keypoints essential for students' comprehension and to generate explanations grounded in those keypoints, thereby delivering more comprehensive educational support and ultimately improving the overall quality of their learning experience.

Mathematical Benchmarks. Current LLMs show strong performance on mathematical problems, making them valuable tools for students (Zhuang et al. 2024; Luo et al. 2025). To evaluate these models, traditional mathematical benchmarks (Cobbe et al. 2021; Hendrycks et al. 2021) have been crucial in assessing reasoning capabilities. With the growing multimodal capabilities of LLMs, benchmarks such as MathVista (Lu et al. 2023), Math-Vision (Wang et al. 2024a), and MathVerse (Zhang et al. 2024) have extended this evaluation to image-based math problems. Recent efforts like OlympiadBench (He et al. 2024) and MM-MATH (Sun et al. 2024) further assess models not just on final answers but also on their reasoning processes. However, most existing benchmarks focus solely on problem-solving, overlooking educational objectives. To address this gap, we introduce ME2, which advances beyond problem-solving to evaluate a model's capacity to generate visually and logically coherent explanations and key visual cues that support effective instructional use.

3 ME2 Benchmark

ME2 is a multimodal solution explanation benchmark consisting of 1,000 instances. Each of which contains a problem text (T_p), a problem image (I_p), an explanatory solution text (T_s), a solution image (I_s), and visual keypoints (VK) that newly highlight elements crucial for understanding (e.g., lines, angles), and a concise explanation summary (T_s^{tldr}) to anchor the model's explanatory solution direction. To create a benchmark that can assess the multimodal solution explanation capabilities of MLLMs, we define the visual keypoint VK and summary of the explanation T_s^{tldr} in Section 3.1. An overview of ME2 and its construction is illustrated in Figure 2.

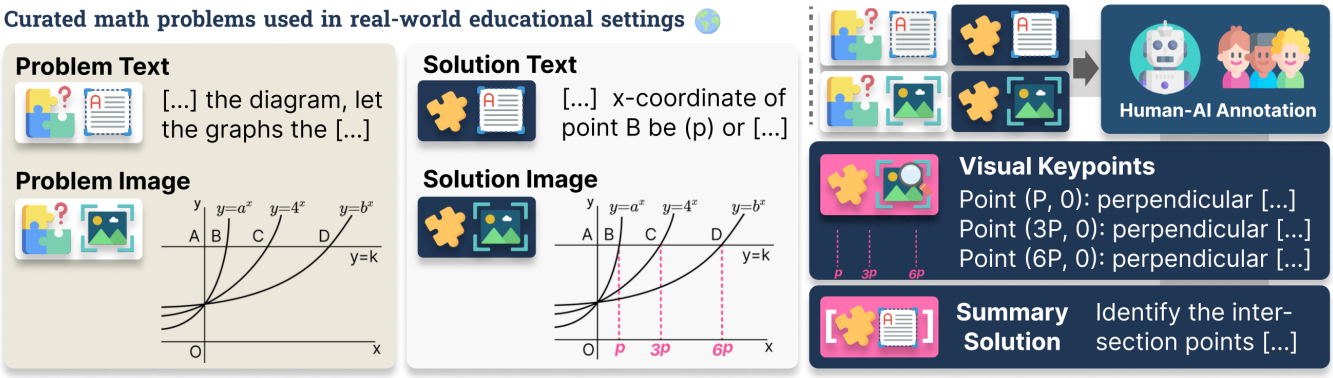

Figure 2: An overview of the ME2 benchmark. The ME2 consists of multimodal problem–solution pairs curated from real-world educational settings, along with visual keypoints and explanation summaries generated through a Human–AI annotation.

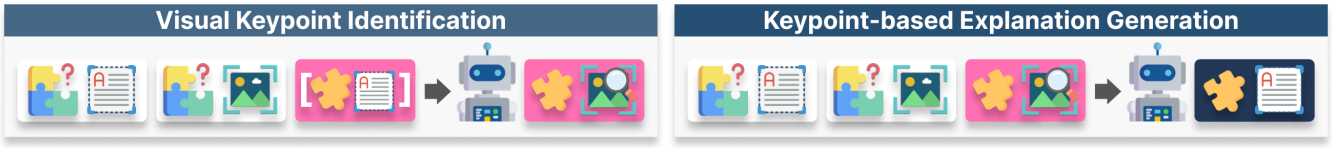

Figure 3: We propose two subtasks to robustly analyze multimodal solution explanation capacity: (1) Visual Keypoint Identification, which challenges machines to recognize visual keypoints useful for subsequent explanation, and (2) Keypoint-based Explanation Generation, which requires models to generate explanations that explicitly reference the identified visual keypoints.

3.1 Benchmark Construction

In-house Data Curation. We extract 1,000 instances of multimodal problem–solution pairs $\langle T_p, I_p, T_s, I_s \rangle$ from an in-house mathematics education platform. All instances were authored by domain experts in mathematics to support effective student learning and are derived from materials authentically used in real-world educational contexts. The instances are written in Korean and span middle- to high-school levels, with a primary focus on geometry and graph theory. All benchmark data were carefully curated to ensure compliance with copyright regulations.

To benchmark the models' ability to recognize visual keypoints, we ensure that each solution image I_s is derived from the corresponding problem image I_p by adding only new elements such as points, angles, lines, regions, and symbols while preserving the original structure. For instance, in Figure 2, points are added to problem image I_p to produce solution image I_s. This setup allows us to accurately evaluate whether a model can identify the critical visual keypoints and effectively incorporate them into its explanation.

We strictly curate the dataset to single-image math problems with either multiple-choice or short-answer formats. We focus on the domains of geometry and graph to ensure that visual context is essential for solving each problem. Each sample in ME2 consists of two natural language texts (the problem text T_p and the solution text T_s) and two RGB images (the problem image I_p and the solution image I_s).

Annotation Process. To create the textual-form visual keypoints, we streamline the simple yet labor-intensive task of comparing problem and solution images by using GPT-4o (Achiam et al. 2023) as an auxiliary tool. The model produces an initial set of keypoints $\{vk_1^{\text{ai}}, \ldots, vk_n^{\text{ai}}\} \in VK^{\text{ai}}$, which four annotators, each holding a bachelor's degree in science or engineering, verify and refine for precision and consistency with our annotation guidelines:

Any element that is newly added or modified, including points, lines, angles, regions, or symbols such as parallel marks, congruence marks, right-angle marks, or length labels, must be recorded in the format {`element`: `description`}, where `element` identifies the visual feature and `description` explains how it is introduced with reference to surrounding features.

Once the verified keypoints are fixed, human annotators and the AI tool jointly generate a brief, keypoint-aligned summary of each solution text (T_s^{tldr}). Since explanations may follow multiple valid paths (see Appendix), this summary anchors a single solution direction during model explanation generation, ensuring an unambiguous consensus set of visual keypoints. Finally, the entire benchmark was translated from Korean to English using an AI tool and then reviewed by two bilingual annotators. From a 10% subset, annotators achieved substantial agreement, with Cohen's κ of 0.84 (Cohen 1960), indicating strong reliability. Consequently, each ME2 instance is represented as $\langle T_p, I_p, T_s, VK, T_s^{\text{tldr}} \rangle$.

3.2 Data Analysis

The ME2 benchmark consists of 1,000 problem–solution pairs: 763 (76.3%) geometry problems and 237 (23.7%) graph problems. Among these, 605 (60.5%) are multiple-choice questions and 395 (39.5%) are short-answer ques-

Total problem–solution pairs	1,000
- Geometry	763
- Multiple-choice questions	464
- Short-answer questions	299
- Graph	237
- Multiple-choice questions	141
- Short-answer questions	96
Average number of VK	3.73
Maximum words in T_p	211
Maximum words in T_s	361
Maximum words in vk_n	45
Maximum words in T_s^{tldr}	198
Average words in T_p	53.1
Average words in T_s	101.4
Average words in vk_n	12.2
Maximum words in T_s^{tldr}	35.8

Table 1: Statistics of the ME2 benchmark, including problem subjects, types, and instance word counts.

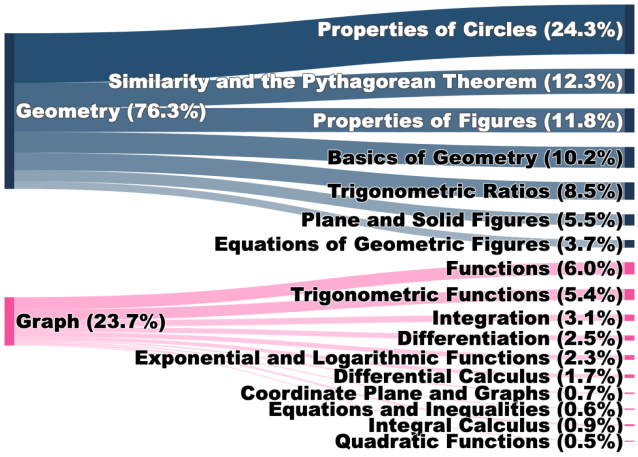

Figure 4: Topic coverage of geometry and graph across 17 chapters in the ME2 benchmark.

tions. It spans 17 chapters (see Figure 4) and 33 sections (see Appendix). On average, each sample contains about 3.8 visual keypoints VK, derived from annotations. These keypoints fall into four main categories: points, lines, regions, and symbols. The symbol category is further divided into parallel marks, equal-length marks, right-angle marks, and length-label marks. Additional statistical details about visual keypoints VK, and length statistics for the problem text T_p, the solution text T_s, and the visual keypoint components vk_n are provided in Table 1.

4 Task Definition

We propose two tasks to evaluate a model's multimodal solution explanation capability, as illustrated in Figure 3. The first task requires (1) identifying visual keypoints useful for subsequent explanation and (2) generating explanations that explicitly reference them. For robust evaluation, we structured the tasks to isolate perceptual and reasoning subskills. Although the design abstracts away some real-world complexity, this two-stage setup still provides a clear and measurable step toward unified, open-ended reasoning.

Visual Keypoint Identification. The first task evaluates the model's ability to identify visual keypoints that are crucial for comprehension. Since a problem may have multiple valid solutions and the corresponding visual keypoints can vary, we provide the model with a solution summary (T_s^{tldr}) that anchors a single explanatory direction. To ensure that the evaluation focuses on keypoint identification rather than problem-solving, the correct answer is also provided. Additionally, because current models cannot reliably generate valid keypoints and open-ended scoring is ambiguous, we adopt a multiple-choice format for robust evaluation.

Given a problem image I_p, its text T_p, the correct answer, and the solution summary T_s^{tldr}, the model must select, from five candidate sets, the visual keypoints (VK) essential for understanding. The four distractor sets are constructed as follows: (1) VK from a problem whose text is semantically similar to T_p; (2) VK from a problem whose solution summary resembles T_s; (3) VK from a problem whose own keypoints closely match the target VK (4) VK from a randomly selected problem. Text similarity was computed using Qwen3 Embedding (Zhang et al. 2025b), and to ensure reliability, human annotators carefully reviewed and corrected any options exhibiting logical inconsistencies.

Keypoint-based Explanation Generation. The second task evaluates whether the model can effectively generate explanatory text grounded in the appropriate visual keypoints. As in the first task, we provide visual keypoints (VK) to guide the model toward a single reasoning path and the correct answer to focus evaluation on keypoint-aligned explanation generation rather than problem-solving.

Given a problem consisting of an image I_p, text T_p, and problem's answer, along with the visual keypoints VK, the model is required to produce a solution explanation T_s that refers to the relevant visual elements.

5 Experiments

Models. We evaluate three categories of MLLMs: (1) **generalist models**, including Molmo 7B (Deitke et al. 2024), LLaVA-1.6 7B (Liu et al. 2024), Qwen2-VL 7B (Wang et al. 2024b), and Qwen2.5-VL 7B & 72B (Bai et al. 2025); (2) **math-specialized models**, including Math-PUMA 7B (Zhuang et al. 2024), URSA 8B (Luo et al. 2025), and Math-LLaVA 13B (Shi et al. 2024); and (3) **proprietary models**, GPT-4o (Achiam et al. 2023) and Gemini 2.0 Flash (Google DeepMind 2024). Details of the experimental setup and prompts are provided in the Appendix.

5.1 Toy: Solution Recognition

The multimodal solution explanation tasks are designed to evaluate specific abilities rather than general problem-solving skills. To examine whether the model genuinely understands how to solve problems, we first perform a preliminary study on ME2 prior to the multimodal explanation task.

		Problem-Solving (Acc)		
Model	Params	Geometric	Graph	Overall
Molmo	7B	0.248	0.194	0.235
LLaVA-1.6	7B	0.147	0.127	0.142
Qwen2-VL	7B	0.274	0.215	0.260
Qwen2.5-VL	7B	0.316	0.224	0.294
Qwen2.5-VL	72B	<u>0.430</u>	**0.300**	<u>0.399</u>
Math-PUMA	7B	0.258	0.194	0.243
URSA	8B	0.055	0.068	0.058
Math-LLaVA	13B	0.202	0.152	0.190
GPT-4o	-	0.274	0.211	0.259
Gemini 2.0 F	-	**0.481**	<u>0.291</u>	**0.436**

Table 2: Experimental results on the *Solution Recognition* toy task from ME2. Models are grouped into three categories: **generalist models** (top), **math-specialized models** (middle), and **proprietary models** (bottom). The best scores are in **bold**, and the second-best scores are <u>underlined</u>.

Metrics. ME2 consists of both multiple-choice and short-answer problems. We report accuracy following the MathVista evaluation protocol (Lu et al. 2023).

Results. Table 2 shows the accuracy of problem-solving. The 7B generalist baseline struggles, while the 72B model performs second best. Somewhat unexpectedly, the math-specialized models perform worse than the generalist models, likely due to hindered instruction-following capabilities. Among the proprietary baselines, GPT-4o struggled similarly to open-source models, whereas Gemini achieved the best performance overall. These results indicate that most MLLMs struggle to recognize the correct solution on ME2, even before performing the multimodal explanation task.

5.2 Visual Keypoint Identification

Metrics. We evaluate baseline performance using accuracy in a multiple-choice setting.

Results. Table 3 summarizes performance on the visual keypoint identification task. The 7B generalist models struggle, while the 72B model remains the second-best performer. In contrast, math-specialized models perform near chance level (Acc = 0.20), indicating severe difficulty in identifying visual cues. Among proprietary models, Gemini achieves the highest performance. Overall, most models struggle to identify visual keypoints even with access to the solution, though proprietary ones perform relatively better.

Since visual keypoint identification was evaluated under the assumption that models can already perform problem-solving, Table 4 reports success rates for each task and for both together. Only 23%, 10%, and 4% of proprietary, generalist, and math-specialized models succeed on both. This result highlights that real educational use remains challenging and that improving problem-solving ability is essential alongside visual keypoint identification.

		VK Identification (Acc)		
Model	Params	Geometric	Graph	Overall
Molmo	7B	0.253	0.312	0.267
LLaVA-1.6	7B	0.260	0.283	0.265
Qwen2-VL	7B	0.273	0.371	0.296
Qwen2.5-VL	7B	0.363	0.532	0.403
Qwen2.5-VL	72B	<u>0.486</u>	<u>0.696</u>	<u>0.536</u>
Math-PUMA	7B	0.194	0.219	0.200
URSA	8B	0.028	0.034	0.029
Math-LLaVA	13B	0.218	0.215	0.217
GPT-4o	-	0.418	0.646	0.472
Gemini 2.0 F	-	**0.529**	**0.726**	**0.576**

Table 3: Experimental results for the *Visual Keypoint Identification* task on ME2, where models are evaluated on their ability to select the correct keypoints from multiple-choices.

Model	PS Only	VKI Only	PS ∩ VKI
Qwen2.5-VL^{7B}	18.9%	29.8%	10.5%
Math-PUMA	19.7%	15.5%	4.6%
Gemini 2.0 F	20.5%	34.5%	23.1%

Table 4: Proportion (%) of cases where each model succeeds only on Problem-Solving (PS), only on Visual Keypoint Identification (VKI), or on both (PS ∩ VKI)

5.3 Keypoint-based Explanation Generation

Metrics. We evaluate the quality of the explanation using three criteria: (1) **Correctness** – whether the model's reasoning is logically sound and leads to a valid solution; (2) **Fidelity** – whether the explanation aligns with the reasoning and intent of the reference, regardless of surface form; (3) **Referencing** – whether the explanation refers to the same key visual components (e.g. points, lines, etc) as the reference. Each criterion is rated on a 5-point Likert scale. We report results from both human evaluators (Zheng et al. 2023) and an LLM-based evaluator using GPT-4o (Achiam et al. 2023). In addition, we report text similarity metrics, including BLEU, ROUGE, METEOR, and BERTScore (Papineni et al. 2002; Lin 2004; Banerjee and Lavie 2005; Zhang et al. 2019), with further details provided in the Appendix.

Results. Table 5 presents the results of the explanation generation task. While most models achieve reasonable Correctness, many fail to follow the intended reasoning path (Fidelity) or reference the given keypoints (Referencing). Generalist models struggle overall, though the Qwen2.5-VL series shows size-dependent improvement. Math-specialized models still fail to produce coherent or instruction-following explanations. In contrast, proprietary models achieve the highest scores, demonstrating stronger abilities in generating well-grounded explanations.

As shown in Table 6, human evaluation exhibits strong correlations with LLM judgments, as indicated by the Spearman coefficients (Zar 2005) (0.770 for Correctness, 0.783

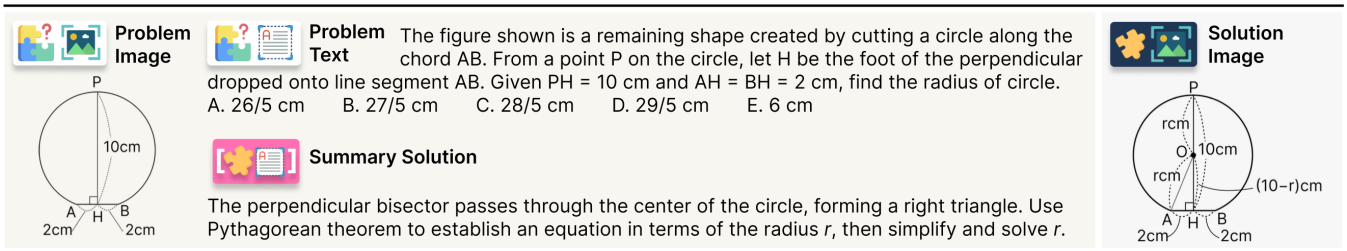

Figure 5: Examples of reasoning processes and final predictions produced by Qwen2.5-VL 7B, Math-PUMA, and Gemini 2.0 Flash on the *Visual Keypoint Identification* task. Qwen2.5-VL demonstrates task understanding and reasoning but produces an incorrect answer, Math-PUMA lacks both, while Gemini 2.0 Flash demonstrates both and produces the correct answer.

Model	Correctness	Fidelity	Referencing
Molmo	2.362	1.906	2.545
LLaVA-1.6	1.815	1.545	1.970
Qwen2-VL	1.844	1.623	1.898
Qwen2.5-VL^{7B}	3.005	2.375	3.132
Qwen2.5-VL^{72B}	3.397	3.048	3.533
Math-PUMA	1.770	1.609	1.581
URSA	1.148	1.140	1.155
Math-LLaVA	2.100	1.360	1.286
GPT-4o	3.784	3.153	3.892
Gemini 2.0 F	**3.849**	**3.489**	**4.103**

Table 5: LLM-based evaluation results for the *Keypoint-based Explanation Generation* task on ME2, rated on a 1-5 Likert scale across three criteria: (1) Correctness, assessing logical validity; (2) Fidelity, measuring alignment with the intent of the reference explanation; and (3) Referencing, evaluating the appropriate use of key visual elements.

Model	Correctness	Fidelity	Referencing
Qwen2.5-VL^{7B}	2.610	2.541	2.610
Math-PUMA	1.041	1.041	1.037
Gemini 2.0 F	**4.423**	**4.171**	**4.256**

Table 6: Human evaluation results for the *Keypoint-based Explanation Generation* task, rated on a 1–5 Likert scale.

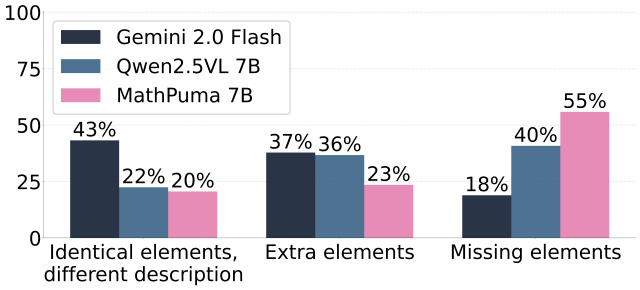

Figure 6: Error analysis for *Visual Keypoint Identification*: (1) correct elements chosen but incorrectly referenced, (2) choices containing more keypoints than required, and (3) choices containing fewer keypoints than needed.

for Fidelity, 0.788 for Referencing; all $p<0.05$). These results show that although most open-source models still struggle, proprietary models and more recent generalist models can generate appropriate explanations.

6 Analyses

6.1 Qualitative Analysis

To analyze how the three categories of models differ in their outputs, we examined the results from representative models in each category: Qwen2.5-VL 7B (generalist), Math-PUMA (specialist), and Gemini 2.0 Flash (proprietary).

Visual Keypoint Identification. Figure 5 shows visual keypoint identification examples from three model categories. Qwen2.5-VL attempts to reason about the most informative keypoints but still selects the wrong option. Math-

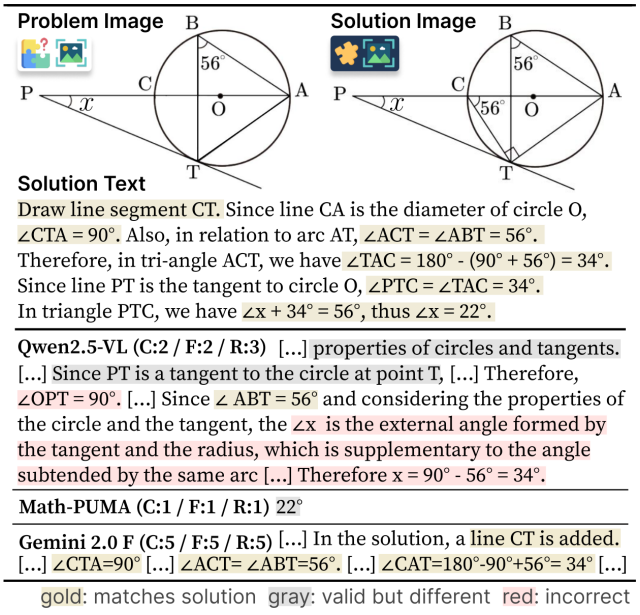

Figure 7: Examples of explanations from three models on the *Keypoint-based Explanation Generation* task, along with their evaluated scores in Correctness (C), Fidelity (F), and Referencing (R). Qwen2.5-VL starts with valid yet unaligned explanations, which soon become incorrect; Math-PUMA generates no explanation; Gemini 2.0 Flash generates solution-aligned explanations.

PUMA shows neither coherent reasoning nor a correct answer. In contrast, Gemini 2.0 Flash correctly interprets the instruction, analyzes the candidates, and chooses the most appropriate keypoints. Overall, similar patterns were consistently observed across examples.

To gain a finer-grained understanding of the models, we analyze their output behaviors on a 10% subset of the dataset, as shown in Figure 6. We categorize all incorrect outputs into three types: (1) electing the correct elements but providing incorrect descriptions, (2) selecting options that contain extra elements, and (3) selecting options with missing required elements. The three models exhibit similar rates for incorrect descriptions and extra elements but differ substantially in missing elements: Math-PUMA shows the highest rate (38%), followed by Qwen2.5-VL (20%). In contrast, Gemini 2.0 Flash achieves the highest accuracy, while Qwen2.5-VL performs moderately, and Math-PUMA remains the least reliable among the three.

Keypoint-based Explanation Generation. Figure 7 presents explanation examples from three model categories. In this example, Qwen2.5-VL scored 2 in Correctness, 2 in Fidelity, and 3 in Referencing. Despite being provided with keypoints, its reasoning diverges from the reference and only partially aligns with the solution image, ultimately leading to an incorrect interpretation. Math-PUMA received a score of 1 across all metrics, as it only returned the final answer without any supporting explanation. In contrast, Gemini scored 5 across all metrics, generating an explanation that mirrored the reference reasoning and consistently referred to the correct key visual elements. As in Figure 6, proprietary models provide coherent explanations, whereas open-source models struggle, with math-specialized ones showing almost no reasoning capability.

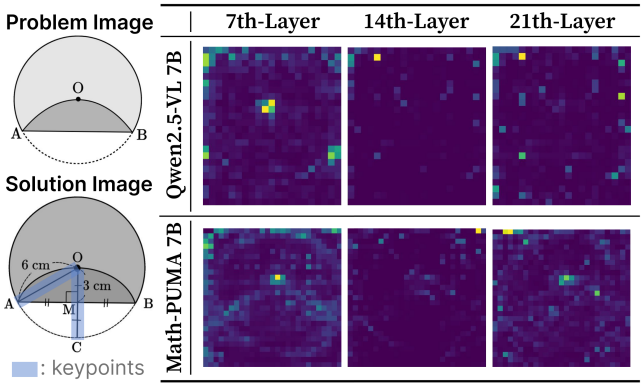

Figure 8: Attention maps from various layers of open-source MLLMs on the *Visual Keypoint Identification* task. While the models attend visually to the problem image, they fail to focus on the keypoints that are most relevant for explanation.

6.2 Do MLLMs Attend to Visual Keypoints?

To analyze how current open-source models fail to visually recognize keypoints, we examined the attention maps of both generalist and math-specialized models on the Visual Keypoint Identification task. Figure 8 illustrates the attention patterns, showing that both models tend to attend well to both global and local regions of the input image. However, despite the explicit inclusion of visual keypoints in the options and summary solution, neither model effectively focuses on the relevant visual regions. A similar pattern is observed in the Keypoint-based Explanation Generation task. While current MLLMs are capable of attending to general visual content (Zhang et al. 2025a), our analysis suggests that they still lack robust mathematical visual grounding ability. Improving visual grounding in mathematical contexts will likely be essential for future models to perform well on the multimodal solution explanation task.

7 Conclusion

We introduce the *multimodal solution explanation* task and the ME2 benchmark, which assess multimodal mathematics-teaching capabilities through two complementary subtasks: identifying essential visual keypoints and generating explanations grounded in those keypoints. Our experimental results demonstrate that current models struggle with both problem-solving and visual keypoint identification, and that this performance gap becomes more pronounced for open-source models in the explanation generation task. Through our analysis, we find that this limitation stems from the lack of math-specific visual grounding and robust visually grounded reasoning. Enhancing these two abilities will be essential for applying MLLMs effectively in educational settings.

Ethics Statement

In this paper, we introduce the ME2 benchmark, curated from an in-house mathematics education platform[1]. All materials were reviewed to ensure full compliance with copyright and data usage regulations. Data annotation was conducted by bilingual annotators holding undergraduate degrees in relevant fields, ensuring adequate mathematical and linguistic proficiency.

Acknowledgments

This work was partly supported by an Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government (MSIT) (No. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University), No.RS-2025-02263598, Development of Self-Evolving Embodied AGI Platform Technology through Real-World Experience), the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00354218, RS-2024-00353125). We express special thanks to KAIT GPU project. The ICT at Seoul National University provides research facilities for this study.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Arcavi, A. 2003. The role of visual representations in the learning of mathematics. *Educational studies in mathematics*, 52(3): 215–241.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Chevalier, A.; Geng, J.; Wettig, A.; Chen, H.; Mizera, S.; Annala, T.; Aragon, M. J.; Fanlo, A. R.; Frieder, S.; Machado, S.; et al. 2024. Language models as science tutors. *arXiv preprint arXiv:2402.11111*.

[1]https://mathpresso.com/en

Clark, J. M.; and Paivio, A. 1991. Dual coding theory and education. *Educational psychology review*, 3: 149–210.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46.

Deitke, M.; Clark, C.; Lee, S.; Tripathi, R.; Yang, Y.; Park, J. S.; Salehi, M.; Muennighoff, N.; Lo, K.; Soldaini, L.; et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Feng, T.; Wang, Z.; and Sun, J. 2023. Citing: Large language models create curriculum for instruction tuning. *arXiv preprint arXiv:2310.02527*.

Google DeepMind. 2024. Introducing Gemini 2.0: our new AI model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/.

He, C.; Luo, R.; Bai, Y.; Hu, S.; Thai, Z. L.; Shen, J.; Hu, J.; Han, X.; Huang, Y.; Zhang, Y.; et al. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Hu, B.; Zheng, L.; Zhu, J.; Ding, L.; Wang, Y.; and Gu, X. 2024. Teaching plan generation and evaluation with GPT-4: Unleashing the potential of LLM in instructional design. *IEEE Transactions on Learning Technologies*.

Jia, J.; Wang, T.; Zhang, Y.; and Wang, G. 2024. The comparison of general tips for mathematical problem solving generated by generative AI with those generated by human teachers. *Asia Pacific Journal of Education*, 44(1): 8–28.

Lee, J.; Park, K.; and Park, J. 2024. VISTA: Visual Integrated System for Tailored Automation in Math Problem Generation Using LLM. arXiv:2411.05423.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, W.; Hu, H.; Zhou, J.; Ding, Y.; Li, J.; Zeng, J.; He, M.; Chen, Q.; Jiang, B.; Zhou, A.; et al. 2023. Mathematical language models: A survey. *arXiv preprint arXiv:2312.07622*.

Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.-W.; Galley, M.; and Gao, J. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Luo, R.; Zheng, Z.; Wang, Y.; Yu, Y.; Ni, X.; Lin, Z.; Zeng, J.; and Yang, Y. 2025. URSA: Understanding and Verifying Chain-of-thought Reasoning in Multimodal Mathematics. *arXiv preprint arXiv:2501.04686*.

Macina, J.; Daheim, N.; Chowdhury, S. P.; Sinha, T.; Kapur, M.; Gurevych, I.; and Sachan, M. 2023. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. *arXiv preprint arXiv:2305.14536*.

Mukul, E.; and Büyüközkan, G. 2023. Digital transformation in education: A systematic review of education 4.0. *Technological forecasting and social change*, 194: 122664.

Paivio, A. 1990. *Mental representations: A dual coding approach*. Oxford university press.

Paivio, A. 2013. *Imagery and verbal processes*. Psychology Press.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Pardos, Z. A.; and Bhandari, S. 2024. ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *Plos one*, 19(5): e0304013.

Park, J.-W.; Park, S.-J.; Won, H.-S.; and Kim, K.-M. 2024. Large Language Models are Students at Various Levels: Zero-shot Question Difficulty Estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 8157–8177.

Shi, W.; Hu, Z.; Bin, Y.; Liu, J.; Yang, Y.; Ng, S.-K.; Bing, L.; and Lee, R. K.-W. 2024. Math-LLaVA: Bootstrapping Mathematical Reasoning for Multimodal Large Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 4663–4680. Miami, Florida, USA: Association for Computational Linguistics.

Stylianou, D. A. 2010. Teachers' conceptions of representation in middle school mathematics. *Journal of mathematics Teacher education*, 13: 325–343.

Sun, K.; Bai, Y.; Qi, J.; Hou, L.; and Li, J. 2024. MM-MATH: Advancing Multimodal Math Evaluation with Process Evaluation and Fine-grained Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 1358–1375.

Uesato, J.; Kushman, N.; Kumar, R.; Song, F.; Siegel, N.; Wang, L.; Creswell, A.; Irving, G.; and Higgins, I. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Wang, K.; Pan, J.; Shi, W.; Lu, Z.; Ren, H.; Zhou, A.; Zhan, M.; and Li, H. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37: 95095–95169.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Zar, J. H. 2005. Spearman rank correlation. *Encyclopedia of Biostatistics*, 7.

Zhang, J.; Khayatkhoei, M.; Chhikara, P.; and Ilievski, F. 2025a. Mllms know where to look: Training-free perception of small visual details with multimodal llms. *arXiv preprint arXiv:2502.17422*.

Zhang, R.; Jiang, D.; Zhang, Y.; Lin, H.; Guo, Z.; Qiu, P.; Zhou, A.; Lu, P.; Chang, K.-W.; Qiao, Y.; et al. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, 169–186. Springer.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhang, Y.; Li, M.; Long, D.; Zhang, X.; Lin, H.; Yang, B.; Xie, P.; Yang, A.; Liu, D.; Lin, J.; Huang, F.; and Zhou, J. 2025b. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.

Zhuang, W.; Huang, X.; Zhang, X.; and Zeng, J. 2024. Math-puma: Progressive upward multimodal alignment to enhance mathematical reasoning. *arXiv preprint arXiv:2408.08640*.