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
jeirfe@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, aside from recent large-scale open-source and closed-source models, most generalist open-source models, and even math-specialist models, struggle with the multimodal solution explanation task. This highlights a significant gap in current LLMs' ability to perform 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.

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

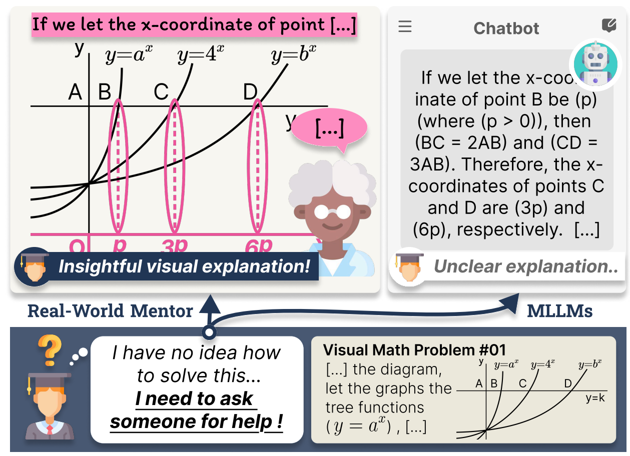

Figure 1: A student solving a math problem often benefits from visual cues, such as lines, symbols, or highlights, that human instructors use to aid understanding, unlike current AI models that focus solely on textual solutions. To serve as effective educational assistants, machines must go beyond answer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

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).

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 opportunities 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,

[*]These authors contributed equally.

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 focus solely on problem-solving and overlook educational objectives, making them insufficient for developing models capable of providing such forms of multimodal instructional support (Hendrycks et al. 2021; Lu et al. 2023; Wang et al. 2024a).

To address these limitations, we introduce *multimodal solution explanation*, a novel task designed to evaluate and advance models' ability to generate educationally effective, visually grounded mathematical explanations. To benchmark performance on multimodal solution explanation, we propose Multimodal Explanations for Mathematics Education (**ME2**) benchmark. The multimodal solution explanation task requires (1) identifying visual keypoints that are not present in the original problem but crucial for understanding (e.g., lines, angles, annotations) and (2) generating explanatory text that explicitly refers to them. The ME2 includes not only the original problem images and text, but also annotations of the visual keypoints that illustrate how the solution images differ from the originals, as well as keypoint-based solution explanations. Notably, we emphasize the educational value of ME2, as it is derived from real-world instructional contexts.

Experiments on ME2 demonstrate that, while closed-source models show potential on multimodal solution explanation, most current open-source generalist models still struggle to reliably identify visual keypoints. Even mathematical models, which are trained on mathematical domains, remain primarily focused on problem-solving and often do not produce sufficiently structured outputs to highlight critical visual details. These findings underscore a critical opportunity to guide future model development toward more effective visually grounded reasoning in educational contexts. We believe that ME2 will catalyze research on strengthening visually grounded mathematical reasoning and advancing multimodal solution explanation models that can serve as effective, 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) highlighting differences from the original, 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.

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 mate-

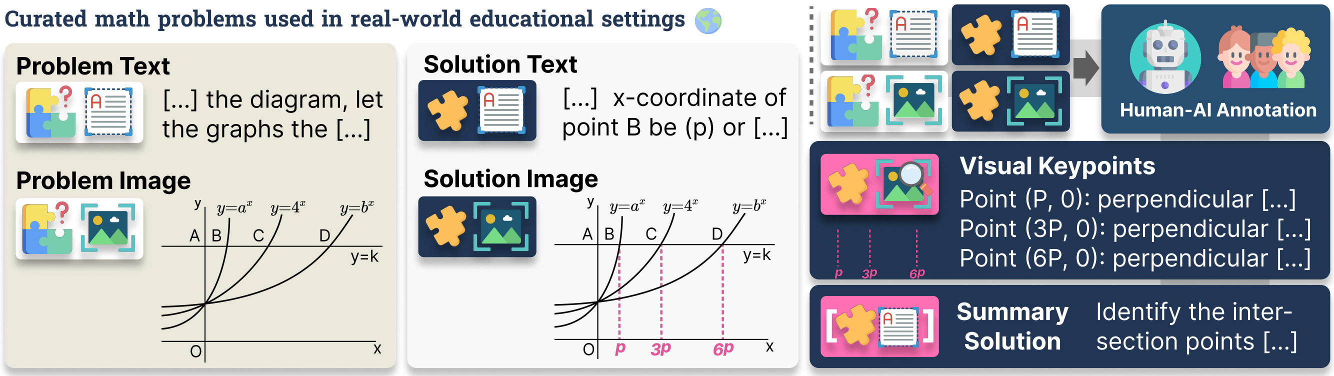

Figure 2: An overview of the ME2 benchmark. The ME2 consists of multimodal problem–solution pairs curated from real-world educational settings, along with visual keypoints and explanation summaries generated through a Human–AI annotation.

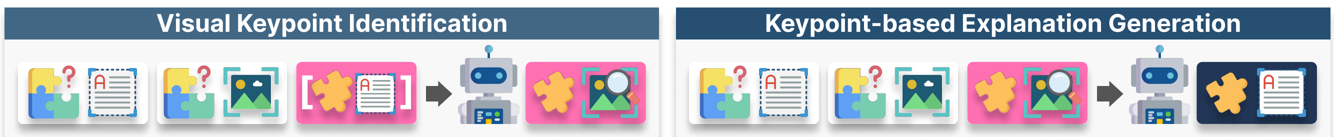

Figure 3: We propose two subtasks to robustly analyze multimodal solution explanation capacity: (1) Visual Keypoint Identification, which challenges machines to recognize visual keypoints useful for subsequent explanation, and (2) Keypoint-based Explanation Generation, which requires models to generate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

rials 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 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 $\{vk_1^{ai}, \dots, vk_n^{ai}\} \in VK^{ai}$, which four annotators, each holding a bachelor's degree in science or engineering, verify and refine for accuracy and compliance with our formatting 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, we generate a brief, keypoint-aligned summary of each solution text (T_s^{tldr}). Since explanations may follow multiple valid paths (see Supplementary Material), this summary anchors a single solution direction during model explanation generation, ensuring an unambiguous consensus set of visual keypoints. As with the keypoints, AI tool produces the initial draft, and human annotators review and refine it for clarity and consistency. 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 questions. 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

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	3.8
Maximum words in T_p	211
Maximum words in T_s	361
Maximum words in vk_n	45
Maximum words in T_s^{tldr}	123
Average words in T_p	53.1
Average words in T_s	102.4
Average words in vk_n	12.2
Maximum words in T_s^{tldr}	35.7

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

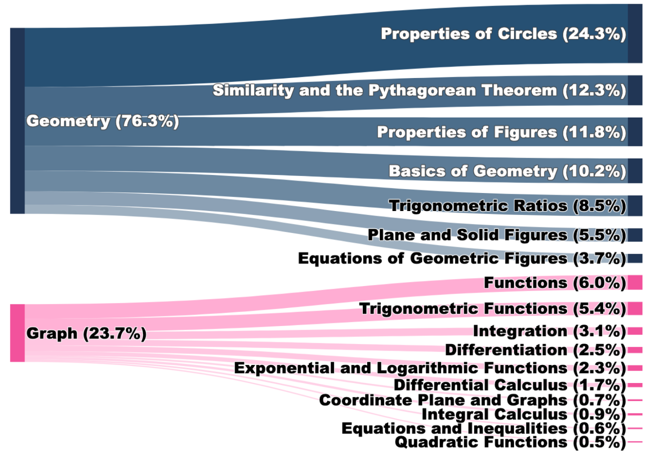

Figure 4: Topic coverage of geometry and graph across 17 chapters in the ME2 benchmark.

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. Multimodal solution explanation requires (1) identifying visual keypoints useful for subsequent explanation and (2) generating explanations that explicitly reference them. These tasks assess how well a model understands multimodal educational elements and generates informative, keypoint-aligned explanations, and they are designed to isolate perceptual and reasoning subskills for controlled evaluation. The two tasks are illustrated in Figure 3.

Visual Keypoint Identification. The first task evaluates the model's ability to identify visual keypoints that are crucial for comprehension. To avoid ambiguity from multiple valid solution paths, we provide the model with a solution summary (T_s^{tldr}) that anchors a single reasoning direction. 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. We compute text similarity using Qwen3 Embedding (Zhang et al. 2025b).

Keypoint-based Explanation Generation. The second task evaluates whether the model can effectively generate explanatory text grounded in the appropriate visual keypoints. We provide the set of visual keypoints (VK) to eliminate ambiguity and guide the model toward a single correct reasoning path. 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.

Although the design abstracts away some real-world complexity, this two-stage structure still offers a clear and measurable step toward unified open-ended reasoning.

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

Inspired by real-world tutoring settings where tutors typically know the solution path, the multimodal solution explanation tasks are designed under the assumption that the model has access to the correct solution, either in the form of a summary solution or a set of visual keypoints. To assess whether the model understands how to solve problems, we first conduct a preliminary study on ME2 before proceeding with the multimodal solution explanation task.

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 each model on the Solution Recognition toy task. The 7B generalist baseline

		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>.

		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)

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 open-source models struggle to recognize the correct solution on ME2, even before performing the multimodal solution 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 difficulty detecting visual cues or weakened instruction-following ability. Among proprietary models, GPT-4o improves noticeably and Gemini achieves the highest performance. Overall, most open-source models struggle to identify visual keypoints even with access to the solution.

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, indicating that real educational use remains challenging and that improving problem-solving ability is essential alongside visual keypoint identification.

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

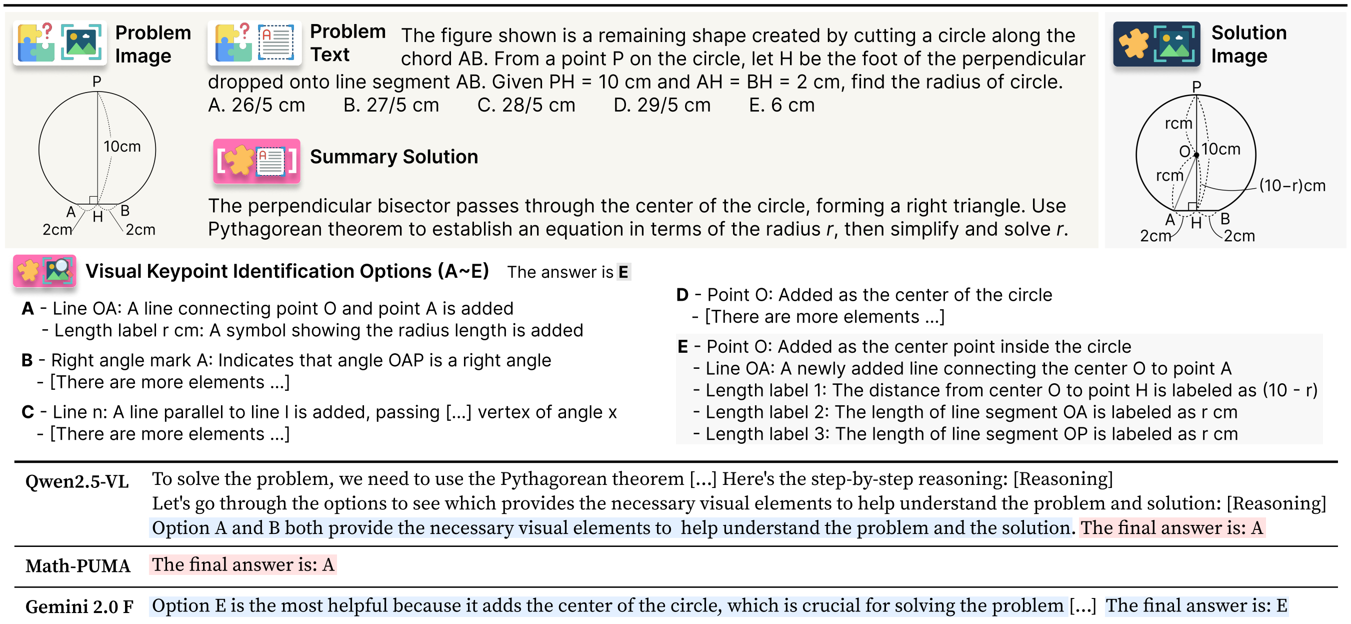

Figure 5: Examples of reasoning processes and final predictions produced by Qwen2.5-VL 7B, Math-PUMA, and Gemini 2.0 Flash on the *Visual Keypoint Identification* task. Qwen2.5-VL demonstrates task understanding and reasoning but produces an incorrect answer, Math-PUMA lacks both, while Gemini 2.0 Flash 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	<u>3.784</u>	<u>3.153</u>	<u>3.892</u>
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.

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-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.

Model	Correctness	Fidelity	Referencing
Qwen2.5-VL^{7B}	<u>2.610</u>	<u>2.541</u>	<u>2.610</u>
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.

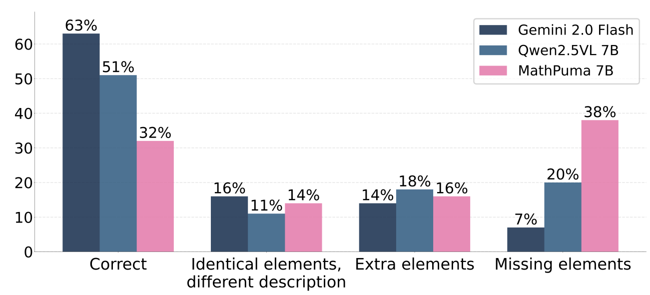

Figure 6: Error analysis for *Visual Keypoint Identification*: (1) Correct choice rates, (2) cases where the correct element was chosen but referenced incorrectly, (3) choices containing more keypoints than required, and (4) choices containing fewer keypoints than needed.

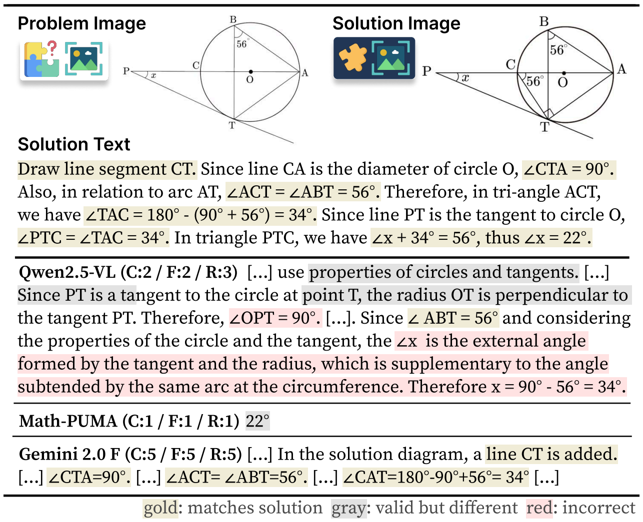

Figure 7: Examples of explanations from three models on the *Keypoint-based Explanation Generation* task, along with their evaluated scores in Correctness (C), Fidelity (F), and Referencing (R). Qwen2.5-VL starts with valid yet unaligned explanations, which soon become incorrect; Math-PUMA generates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

To more precisely compare the models, we analyzed failure cases on a 10% subset of the dataset, as shown in Figure 6. We categorize errors into three types – electing the correct elements but providing incorrect descriptions, selecting options that contain extra elements, and 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 and the most stable error pattern, whereas 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 for Correctness, 2 for Fidelity, and 3 for Referencing. Although the model is given keypoints, its approach is not aligned with the reference solution. Its explanation references visual cues such as angles, lengths, symbols, and triangles, which indicates partial alignment with the solution image. However, it ultimately misinterprets the problem and produces incorrect reasoning. 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 in all metrics. It generated an explanation as if the keypoints were embedded in the solution image, following the same reasoning structure as the reference and correctly referring to the same key visual elements throughout the solution.

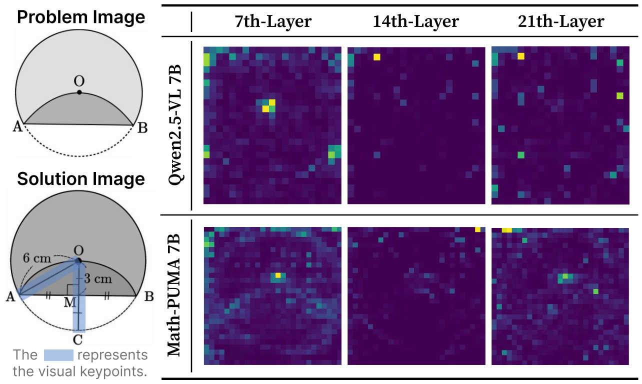

Figure 8: Attention maps from various layers of open-source MLLMs on the *Visual Keypoint Identification* task. While the models attend visually to the problem image, they fail to focus on the keypoints that are most relevant for explanation.

6.2 Do MLLMs Attend to Visual Keypoints?

To analyze whether current open-source models can accurately identify visual keypoints in images, we examined the attention maps of generalist and math-specialized models on the Visual Keypoint Identification task. Figure 8 illustrates the attention patterns of the generalist and math-specialized models. Both models show a tendency to attend 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.

These findings suggest that current MLLMs are capable of attending to visual content, as also shown in prior work (Zhang et al. 2025a) demonstrating that models can reliably focus on general, familiar objects. However, their ability to visually ground information in unfamiliar domains, such as mathematics, remains limited. 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 MLLMs struggle with both problem solving and visual keypoint identification in ME2, with reliable performance emerging only from proprietary models and the latest open-source generalist models. Math-specialized models perform even more poorly, and this performance gap becomes more pronounced in the explanation generation task. These findings underscore the need for more advanced math-specific visual grounding, along with stronger visually grounded reasoning and solution-path tracking, to enable effective multimodal mathematical explanations.

8 Ethical Considerations

In this paper, we introduce the ME2 benchmark, curated from an in-house mathematics education platform. All materials were reviewed to ensure compliance with copyright regulations. Data annotation was performed by bilingual annotators with undergraduate degrees, ensuring sufficient mathematical problem-solving skills.

9 Acknowledgement

This was supported by Digital Appliances (DA) Business, Samsung Electronics Co.,Ltd and National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) (Nos. RS-2024-00354218 and RS-2024-00353125) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)).

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*.

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*.

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

Technical Appendix

A Overview

This supplementary material includes the following sections: (1) detailed experimental settings used in our benchmark, (2) a description of our benchmark, (3) prompts used for model response generation, and (4) ethical considerations.

B Experiment Details

In this study, experiments were conducted using the LMMs-eval repository [1]. This repository provides a comprehensive framework for evaluating multi-modal models across various tasks.

B.1 Computational Resources

For closed-source models, we used the OpenAI API and Gemini Developer API to infer the output of GPT-4o and Gemini 2.0. For open-source models such as Math-PUMA, URSA, MathLLaVA, LLaVA, QWEN-2-VL, QWEN-2.5-VL, and Molmo, inference was performed using a NVIDIA A6000 48GB GPU. While the exact inference speed varies depending on the task, model, and lengths of prompts and responses, a query takes about 50 seconds to be answered.

B.2 Evaluated Models

When conducting experiments with open-source multimodal models, we leveraged the official implementation codes in conjunction with publicly available weights from the Huggingface Hub [2]. The following model parameters were used for each model:

- **Math-PUMA:** `Math-PUMA/Math-PUMA_Qwen2VL-7B`
- **URSA:** `URSA-MATH/URSA-RM-8B`
- **Math-LLaVA:** `Zhiqiang007/Math-LLaVA`
- **llava-1.6:** `llava-hf/llava-v1.6-mistral-7b-hf`
- **Qwen2-VL-7B:** `Qwen/Qwen2-VL-7B-Instruct`
- **Qwen2.5-VL-7B:** `Qwen/Qwen2.5-VL-7B-Instruct`
- **Qwen2.5-VL-72B:** `Qwen/Qwen2.5-VL-72B-Instruct`
- **Molmo:** `allenai/Molmo-7B-D-0924`

These models were evaluated in our benchmark, which included tasks designed to assess both visual understanding and textual explanation capabilities. The selection of models spans a range of architectures and performance levels, providing insights into current advancements in multi-modal learning.

[1]https://github.com/EvolvingLMMs-Lab/lmms-eval

[2]https://huggingface.co/models

B.3 LLM Evaluation

In all LLM-based evaluations, we used the `gpt-4o-2024-08-06` endpoint. For the Keypoint-based Explanation Generation task, we compared the rankings obtained when Math-PUMA and GPT outputs were evaluated separately by Gemini and GPT. The resulting Kendall's values were 0.90 for Correctness, 0.84 for Fidelity, and 0.92 for Referencing. Although evaluation bias is often a concern when an LLM assesses models from the same family, the high agreement between the GPT-judge and Gemini-judge indicates that no substantial bias is present.

B.4 Human Evaluation

Three evaluators, all holding a bachelor's or master's degree in engineering, are assessing AI model outputs for 80 problems. Specifically, they are evaluating the results produced by the Math-PUMA, Qwen2.5-VL, and Gemini 2.0 Flash models, which represent the math-specialized, generalist, and proprietary models categories. Each criterion is being rated on a five-point Likert scale. LLM–human agreement shows strong correlations, as indicated by the Spearman coefficients (0.770 for Correctness, 0.783 for Fidelity, and 0.788 for Referencing; all $p < 0.05$). Human–Human agreement, measured by Krippendorff's α, reached 0.696 for Correctness, 0.571 for Fidelity, and 0.612 for Referencing.

B.5 Automatic Metrics Evaluation

We report additional evaluation results for Keypoint-based Explanation Generation task using automatic metrics, including BLEU, ROUGE, METEOR, and BERTScore. The detaileed scores can be found in Table A.

C Benchmark Details

Our ME2 benchmark consists of a total of 17 chapters and 33 sections as shown in Table B.

D Prompts

This section compiles all the prompts used in our experiments. The prompts shown in Figure A, Figure B, and Figure C are used to generate model outputs for the Solution Recognition toy task, Visual Keypoint Identification, and Keypoint-based Explanation Generation tasks, respectively. For the Keypoint-based Explanation Generation task, model responses are evaluated using the prompts in Figure D.

Model	Params	BLEU-2	BLEU-4	ROUGE-L	METEOR	BERTScore
Molmo	7B	<u>0.158</u>	0.067	0.187	0.310	<u>0.842</u>
LLaVA-1.6	7B	0.130	0.059	0.176	0.287	0.835
Qwen2-VL	7B	**0.176**	**0.087**	**0.237**	0.288	**0.854**
Qwen2.5-VL	7B	0.099	0.038	0.193	0.284	0.819
Qwen2.5-VL	72B	0.097	0.043	0.190	<u>0.316</u>	0.821
Math-PUMA	7B	0.006	0.002	0.119	0.058	0.818
URSA	8B	0.020	0.006	0.079	0.075	0.735
Math-LLaVA	13B	0.112	0.057	0.147	0.252	0.817
Gemini 2.0 Flash	-	0.149	<u>0.083</u>	<u>0.210</u>	**0.367**	<u>0.842</u>
GPT-4o	-	0.095	0.045	0.161	0.301	0.815

Table A: Experimental results of automated evaluation metrics for the *Keypoint-based Explanation Generation* task on ME2.

Chapter Title	**Section Title**
Basics of Geometry	Basic Geometric Figures Construction and Congruence
Coordinate Plane and Graphs	Coordinate Plane and Graphs
Differential Calculus	Differentiation of Various Functions
Differentiation	Derivative and Derivative Function Applications of Derivatives
Equations and Inequalities	Quadratic Equations and Functions
Equations of Geometric Figures	Transformations of Figures Equation of a Circle Coordinate Plane Equations of Straight Lines
Exponential and Logarithmic Functions	Exponential and Logarithmic Functions
Functions	Linear Functions and Their Graphs Functions Rational and Irrational Functions Relationship Between Linear Functions and Equations
Integral Calculus	Applications of Definite Integrals Various Integration Techniques
Integration	Indefinite and Definite Integrals Applications of Definite Integrals
Plane and Solid Figures	Properties of Solid Figures Properties of Plane Figures
Properties of Circles	Circle and Line Inscribed Angles
Properties of Figures	Properties of Quadrilaterals Properties of Triangles
Quadratic Functions	Graph of the Quadratic Function $y = ax^2 + bx + c$
Similarity and the Pythagorean Theorem	Pythagorean Theorem Similarity of Figures Applications of Similarity
Trigonometric Functions	Meaning and Graphs of Trigonometric Functions Law of Sines and Law of Cosines
Trigonometric Ratios	Trigonometric Ratios Applications of Trigonometric Ratios

Table B: Overview of the 17 chapters and their corresponding 33 sections covered in the dataset.

You should choose a set of visual elements from the multiple-choice options (A, B, C, D, or E) that best reflect how a teacher would visually guide a student to understand and solve the problem.

Problem:
As shown in the figure, there are 5 points: A, B, C, D, and E. When selecting two points among them to form straight lines and rays, let the number of straight lines be a and rays be b. Find the value of a + b.

Answer: **19**
The solution process for the problem is as follows:

Count the possible straight lines formed by selecting pairs of points, then count the rays formed by considering directionality. Add both counts to find the total.

A.- line AE: A line connecting point A and E

B.- Symbol a: Represents the line extending from the upper left to the lower right
- Symbol b: Represents the line extending from the lower left to the upper right
- Symbol c: Represents the horizontal line

C.- Line AB: A line extended from side AB of the hexagon
- Line BC: A line extended from side BC of the hexagon
- Line CD: A line extended from side CD of the hexagon
- Line DE: A line extended from side DE of the hexagon
- Line EF: A line extended from side EF of the hexagon
- Line FA: A line extended from side FA of the hexagon

D.- Auxiliary line BD: A line segment connecting point B and point D is added

E.- Line AE: A straight line connecting points A and E
- Line BE: A straight line connecting points B and E
- Line CE: A straight line connecting points C and E
- Line DE: A straight line connecting points D and E

Based on this reasoning guidance, select **only one** of the option (A, B, C, D, or E) whose visual elements would be most helpful for students in understanding the problem and its solution.

Think carefully about how the selected visual elements support the reasoning process. You may briefly explain your thinking, but your response **must end** with the following format:
The final answer is: A, B, C, D, or E
IMPORTANT!! Your final response must END with the format.

Figure A: Prompt used for the *Visual Keypoint Identification* task in the ME2 benchmark. Prompt inputs are **boldfaced**.

Q: **As shown in the figure, there are 5 points: A, B, C, D, and E. When selecting two points among them to form straight lines and rays, let the number of straight lines be a and rays be b. Find the value of a + b.**
Answer: **19**

Difference between the original image and the solution image
Line AE: A straight line connecting points A and E
Line BE: A straight line connecting points B and E
Line CE: A straight line connecting points C and E
Line DE: A straight line connecting points D and E

You are a math teacher helping students understand how to solve problems clearly and effectively.
Given a problem description, problem image and a list of key elements introduced or highlighted in the solution image, write an educational explanation that helps students.
Additionally, this problem is a problem of **Functions/Linear Functions and Their Graphs** chapter. You should explain the problem in the context of the chapter and section.
Make sure to reference both the original components from the problem image and any new annotations, highlights, or added elements from the solution image to enhance understanding.

OUTPUT Example:
{
  solution_text:
}

Figure B: Prompt used for the *Keypoint-based Explanation Generation* task in the ME2 benchmark. It is designed to generate educationally effective explanations for the given math problem. Prompt inputs are **boldfaced**.

You are a math solver. For the problem below, **your task is ONLY to output the final answer** in one line.
Do NOT provide any explanation, steps, or clarification. Just write the answer.

Problem: **As shown in the figure, there are 5 points: A, B, C, D, and E. When selecting two points among them to form straight lines and rays, let the number of straight lines be a and rays be b. Find the value of a + b.**

Again, only return the final answer. Any additinal text will be considered incorrect.

Figure C: Prompt used for the *Solution Recognition* toy task in the ME2 benchmark. It is designed to generate an answer to the given math problem. Prompt inputs are **boldfaced**.

You are evaluating the quality of an AI-generated explanation for a math problem involving geometry or graph-based reasoning.

You will be given two texts:

1. A reference explanation written by a human teacher.
2. An AI-generated explanation written by a model.

Your task is to compare the two explanations and assess how accurately and effectively the AI-generated explanation captures the key geometric concepts and reasoning presented in the reference.

Please evaluate the model's explanation and provide four scores based on the criteria below:

—
Scoring Criteria

1. Correctness
- Does the reasoning presented by the model make sense and help solve the problem appropriately?
- Rate on a Likert scale: **1, 2, 3, 4, or 5**

2. Reference Alignment
- Does the model follow the same logical reasoning and intent as the reference explanation, even if the wording differs?
- Rate on a Likert scale: **1, 2, 3, 4, or 5**

3. Use of Key Visual Elements
- Does the AI explanation refer to the same critical visual components (e.g., points, lines, angles, shapes) as the reference?
- Alternative terminology is acceptable if it clearly refers to the same element or serves the same purpose.
- Rate on a Likert scale: **1, 2, 3, 4, or 5**

—
Output Format

Important: Report your rating using the exact format below:

Rating: [[x, y, z]]
— where ‘x‘ is your score for correctness, ‘y‘ for reference alignment, and ‘z‘ for use of visual elements.

Figure D: GPT evaluation prompt used to assess model outputs for the *Keypoint-based Explanation Generation* task in ME2. Prompt inputs are **boldfaced**.