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

Technical Appendix

A Experiment Details

In this study, experiments were conducted using the LMMs-eval repository [2]. This repository provides a comprehensive framework for evaluating multi-modal models across various tasks.

A.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.

A.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 [3]. 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.

A.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 τ 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.

A.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.

A.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.

B Benchmark Details

Our ME2 benchmark consists of a total of 17 chapters and 33 sections as shown in Table B. During dataset validation, we also reviewed the options related to visual keypoint identification and confirmed their consistency and reliability.

C Ablation on Solution Summary Anchoring

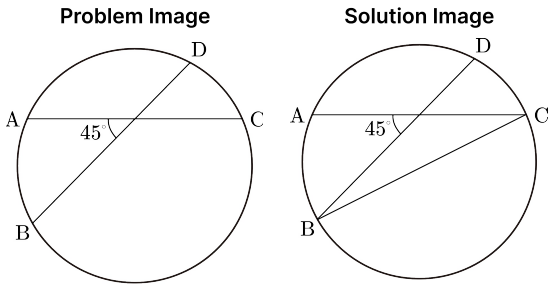
Since a problem may have multiple valid solutions and the corresponding visual keypoints can vary, we provide the model with a solution summary (T_s^{tldr}) that anchors a single explanatory direction. To examine the effect of this anchoring, we analyze the qualitative differences when the summary is not provided in Figure A. As shown, without anchoring, the model's explanations often drift toward alternative reasoning paths or focus on irrelevant keypoints.

D Prompts

This section compiles all the prompts used in our experiments. The prompts shown in Figure B, Figure C, and Figure D 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 E.

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

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



Visual Keypoints

```
[
  {'element': 'Line BC': 'description': 'A line connecting points B and C is added'}
]
```

(a) without solution summary anchoring

```
[
  {"element": "Point O", "description": "The center of the circle"},
  {"element": "Angle AOC", "description": "Central angle subtended by arc AC"},
  {"element": "Angle BOD", "description": "Central angle subtended by arc BD"},
  {"element": "Angle AOB", "description": "Central angle subtended by arc AB"},
  {"element": "Angle COD", "description": "Central angle subtended by arc CD"},
  {"element": "Length label", "description": "r, radius of the circle"},
  {"element": "Line OA", "description": "Line connecting center O to point A"},
  {"element": "Line OB", "description": "Line connecting center O to point B"},
  {"element": "Line OC", "description": "Line connecting center O to point C"},
  {"element": "Line OD", "description": "Line connecting center O to point D"}
]
```

(b) with solution summary anchoring

```
[
  {"element": "Line BC", "description": "Auxiliary line connecting point B and point C"},
  {"element": "Angle BAC", "description": "Inscribed angle BAC"},
  {"element": "Angle BCD", "description": "Inscribed angle BCD"},
  {"element": "Angle BAC", "description": "Angle 1"},
  {"element": "Angle BCD", "description": "Angle 2"}
]
```

Figure A: Comparison of explanations generated with and without the solution summary (T_s^{tldr}). Without anchoring, models often drift to alternative reasoning paths or irrelevant keypoints. (a) Without T_s^{tldr}; (b) With T_s^{tldr}.

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