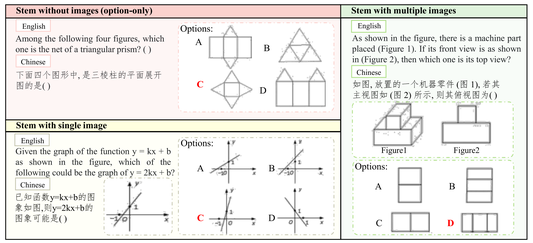
VisioMath: Benchmarking Figure-based Mathematical Reasoning in LMMs

Can Li · Ying Liu · Ting Zhang † · Mei Wang · Hua Huang
School of Artificial Intelligence, Beijing Normal University · Beijing Key Laboratory of Artificial Intelligence for Education · Engineering Research Center of Intelligent Technology and Educational Application, Ministry of Education

1. Problem Setting

Why this benchmark?

- Existing multimodal math benchmarks mostly use a single supporting image and rarely test answer options that are themselves diagrams.
- VisioMath targets fine-grained comparative reasoning across multiple visually similar candidates in realistic K–12 exam settings.
- All answer choices are images, so solving requires cross-figure comparison rather than text-only elimination.

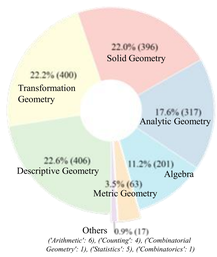


Representative examples: stem-only options, single-image stems, and multi-image stems.

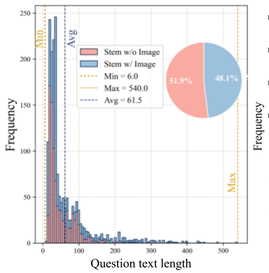
2. Benchmark Overview

1,800	8,070	51	4.48
problems	diagram images	evaluated LMMs	images / problem

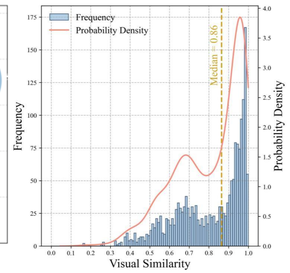
- Collected from Chinese middle- and high-school entrance examinations from 2002–2023.
- Balanced answer distribution across A–D (roughly 24–26%) to reduce positional bias.
- Question text averages 61.5 tokens; option images are often highly similar (median similarity 0.86).



(a) Distribution of subjects.



(b) Distribution of question text length.



(c) Distribution of visual similarity.

Key design choice

The benchmark explicitly couples textual stems with multiple candidate diagrams, making multi-image alignment a first-class challenge rather than a side effect of prompting.

3. Main Results

Even strong LMMs struggle: GPT-4.1= 52.6%, best open-source = 66.9%, best closed-source = 80.9%.
Accuracy consistently drops as inter-image similarity increases, exposing a weakness in comparative visual reasoning.

Models \Image similarity	Avg	[0.16,0.68]	(0.68,0.87]	(0.87,0.96]	(0.96,1]
Human	91.3	95.7	91.2	87.6	89.0
Random	25.6	23.6	24.4	27.8	27.1
Closed-source LMMs					
QwenVL-max (Bai et al., 2023)	44.1	47.3	50.2	41.3	37.6
GPT-4.1 (OpenAI, 2025)	52.6	65.8	56.4	42.9	45.1
Seed1.6-Thinking (ByteDance, 2024)	72.3	82.4	74.2	66.2	66.4
Gemini 2.5 Pro (Comanici et al., 2025)	**80.9**	**86.2**	**83.8**	**76.7**	**76.9**
Open-source LMMs (multi-image input)					
InternVL2.5-2B (Chen et al., 2024a)	24.6	24.2	28.9	22.7	22.7
Qwen2.5-VL-3B-instruct (Bai et al., 2025)	25.4	26.7	27.6	24.4	22.9
R1-Onevison-7B (Yang et al., 2025)	29.6	21.9	32.2	28.9	11.6
Qwen2.5-VL-7B-instruct (Bai et al., 2025)	32.7	33.6	37.8	29.8	29.6
Gemma3-27B (Team, 2025b)	35.3	43.3	41.2	29.6	26.4
Vision-R1-7B (Huang et al., 2025)	36.7	46.7	38.9	30.4	30.9
Qwen2.5-VL-72B-instruct (Bai et al., 2025)	43.7	47.1	50.8	38.0	38.7
GLM-4.5V (Team et al., 2025)	**53.7**	**68.7**	**59.3**	**44.2**	**44.7**
Open-source LMMs (math-oriented)					
MM-PRM-8B (Du et al., 2025)	31.7	37.6	37.1	26.9	25.1
MM-Eureka-7B (Meng et al., 2025)	37.9	45.6	44.0	29.1	**33.1**
MM-Eureka-7B-CPGD (Liu et al., 2025)	**39.4**	**47.8**	**46.0**	**30.9**	32.9

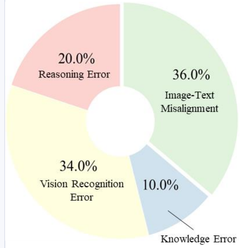
As similarity rises from low to high quartiles, model accuracy decreases across families, indicating that the benchmark stresses fine-grained discrimination rather than broad scene understanding.

Humans plateau after similarity passes a threshold, suggesting later errors are more conceptual than perceptual. LMMs, by contrast, still miss distinctions that humans rarely confuse - pointing to weak visual-text alignment rather than shallow reasoning alone.

Models \GT position	Avg	Question stem w/o images				Question stem with images				
		Avg	A	B	C	D	Avg	A	B	C	D
Human	91.3	92.3	92.5	95.6	93.8	88.5	89.7	94.4	87.6	87.5	88.0
Random	25.6	25.4	24.0	25.6	23.0	28.6	26.0	22.8	27.6	28.4	25.6
Closed-source LMMs											
QwenVL-max (Bai et al., 2023)	44.1	53.4	35.2	62.6	62.5	50.2	34.1	31.1	34.1	32.8	38.6
GPT4.1 (OpenAI, 2025)	52.6	61.6	72.4	59.9	60.2	56.1	42.8	54.8	39.3	43.7	31.9
Seed1.6-Thinking (ByteDance, 2024)	72.3	85.7	**90.3**	**87.2**	82.4	83.9	58.0	71.8	53.7	44.6	59.4
Gemini 2.5 Pro (Comanici et al., 2025)	**80.9**	**86.2**	89.2	84.6	**85.2**	**86.3**	**75.2**	**78.8**	**77.6**	**75.0**	**68.6**
Open-source LMMs (multi-image input)											
InternVL2.5-2B (Chen et al., 2024a)	24.6	27.1	12.8	25.5	36.3	30.2	21.9	10.3	26.2	**38.2**	15.0
Qwen2.5-VL-3B-instruct (Bai et al., 2025)	25.4	26.1	51.0	40.5	14.5	5.9	24.7	18.3	**70.1**	5.4	4.3
R1-Onevison-7B (Yang et al., 2025)	29.6	35.0	38.8	37.4	34.8	30.2	23.7	22.0	32.2	28.9	11.6
Qwen2.5-VL-7B-instruct (Bai et al., 2025)	32.7	39.5	30.1	58.1	39.8	29.8	25.3	8.7	28.5	32.4	34.3
Gemma3-27B (Team, 2025b)	35.3	43.7	67.9	40.1	33.6	38.4	26.2	40.2	24.8	12.3	25.1
Vision-R1-7B (Huang et al., 2025)	36.7	43.7	47.4	57.3	38.7	33.7	29.2	24.5	52.3	29.4	10.6
Qwen2.5-VL-72B-instruct (Bai et al., 2025)	43.7	53.5	36.2	63.9	61.3	49.8	33.0	29.9	37.8	29.9	**35.2**
GLM-4.5V (Team et al., 2025)	**53.7**	**69.1**	**71.9**	**75.8**	**68.4**	**61.2**	**37.2**	**46.5**	42.5	31.4	26.6
Open-Source LMMs (math-oriented)											
MM-PRM-8B (Du et al., 2025)	31.7	38.4	28.1	43.2	44.9	35.7	24.4	10.8	**41.6**	35.3	11.6
MM-Eureka-7B (Meng et al., 2025)	37.9	50.9	**36.2**	**62.1**	52.7	50.1	24.0	**21.1**	22.4	27.4	**25.6**
MM-Eureka-7B-CPGD (Liu et al., 2025)	**39.3**	**51.0**	33.2	54.2	**61.3**	**51.4**	**26.9**	16.2	29.9	**39.7**	23.7

Most LMMs perform worse when the question stem itself contains an image. Once both the stem and the options are visual, the model must integrate multiple visual contexts at once - a clear bottleneck for holistic figure understanding.

4. Error Analysis

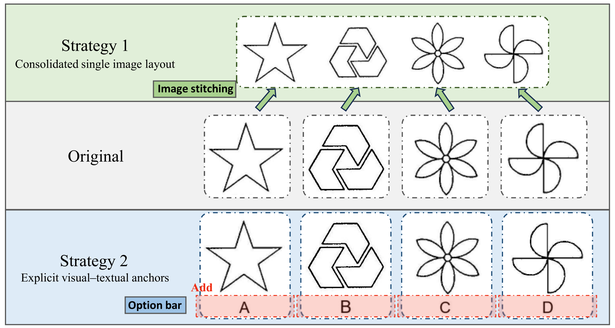


Error composition and the effect of shuffling.

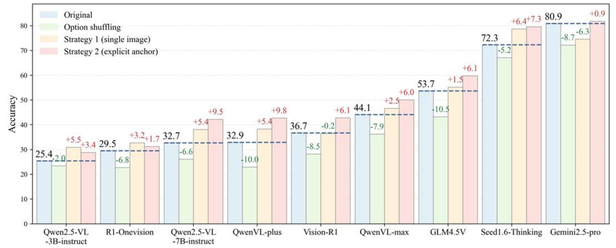
36% of errors are image-text misalignment errors - the single largest category.

This indicates that current failures often emerge before deep reasoning starts: the model first loses track of which visual element corresponds to which textual cue.

5. Alignment-Oriented Strategies



Training-free strategies: single-image stitching and explicit visual–text anchors.



Model	Original	Shuffling	Strategy 1	Strategy 2	Strategy 3
Qwen2.5-VL-3B-instruct (Bai et al., 2025)	25.4	23.4 (-2.0)	30.9 (+5.5)	28.8 (+3.4)	38.0 (+12.6)
Qwen2.5-VL-7B-instruct (Bai et al., 2025)	32.7	26.1 (-6.6)	38.1 (+5.4)	42.5 (+9.8)	43.3 (+10.6)
Qwen2.5-VL-72B-instruct (Bai et al., 2025)	43.7	35.6 (-8.1)	47.6 (+3.9)	50.1 (+6.4)	51.4 (+7.7)
InternVL2.5-2B (Chen et al., 2024a)	24.6	23.1 (-1.5)	26.3 (+1.7)	27.2 (+2.6)	32.2 (+7.6)

- **Shuffling**: Option shuffle tests whether the model truly matches the question stem to the correct option rather than relying on option order or positional shortcuts; performance drops after shuffling indicate weak option-image alignment.
- S1 (single stitched image) reduces the burden of cross-image attention.
- S2 (explicit anchors) makes option-image binding easier.
- S3 (alignment-oriented CoT finetuning) gives the largest gains, up to +12.6 points.

6. Takeaways & QR

Takeaways

- VisioMath is the first benchmark centered on diagrammatic answer options in K–12 mathematics.
- Current LMMs remain brittle when they must align several similar images with textual choices.
- Simple structural changes help, but explicit alignment data is the most effective remedy.
- The benchmark provides a rigorous testbed for educational AI and diagram reasoning.

Scan for code & dataset

github.com/Nefefilibata/VisioMath