PICTUREPROOF: EVALUATING MATHEMATICAL REASONING IN VISION-LANGUAGE MODELS WITH VISUAL PROOFS

Anonymous authors
Paper under double-blind review

ABSTRACT

Human mathematical reasoning is profoundly visual at the level of intuition, discovery, and communication. However, existing visual math benchmarks evaluate on shallow calculation-based tasks rather than deeper modes of reasoning. To address this benchmarking gap, we introduce PICTUREPROOF, an open-ended benchmark of visual mathematical reasoning based on translating visual math proofs into natural language. Because the grading for this task is inherently qualitative, we implement a framework in which VLM outputs are graded at scale with an LLM-judge using human expert-crafted rubrics for each benchmark item. Our evaluations reveal dramatic failures in visual mathematical reasoning both on open-source as well as frontier VLMs. We further demonstrate that some of the failures are caused by a decay in attention to visual tokens during reasoning. We open-source our benchmark—specifically, the visual math proofs in TikZ code format, rubrics, and LLM-judge prompts—in order for the community to run further experiments on mathematical reasoning in VLMs.

1 INTRODUCTION

The relationship between mathematics and visual understanding has a long tradition dating back to the era of Pythagoras and Euclid (Zimmermann & Cunningham, 1991; Palais, 1999). These forms of mathematical visual understanding may range in abstraction level, from informal intuition to concrete diagrams. But across these abstraction levels, visualizations are broadly helpful for revealing connections and managing information within complex mathematical objects. As such, practitioners of mathematics from schoolchildren to research mathematicians use visualizations in their work (Hadamard, 1945; Pólya, 1957).

Although visual math reasoning in vision language models (VLMs) has been studied through an array of visual math benchmarks (Qiao et al., 2024), (Lu et al., 2024), (Zhang et al., 2024), all existing benchmarks rely on easily-verifiable, computation-based tasks, thereby failing to capture the theoretical and intuitive dimension of visual math reasoning. In particular, these benchmark tasks rely on information that is conveyed in diagrams, but they do not require substantial multi-step visual information comprehension and manipulation beyond reading values from the figure.

Beyond the field of mathematics, establishing a challenging visual reasoning task is broadly applicable for studying persistent failure modes in VLMs. A recent body of work has demonstrated failures in VLM visualization comprehension tasks that require deeper levels of reasoning, notably in charts and plots as well as counting (Pandey & Ottley, 2025; Verma et al., 2024; Che et al., 2026). Furthermore, several studies describe a phenomenon in which visual information in VLM's is lost at the language-generation stage, during which language outputs overwhelm visual information (Tartaglini et al., 2025; Liu et al., 2026). Previous VLM studies drew on the unique characteristics of their problem domains in order to probe the nature of these reasoning failures. For this reason, it is useful to develop new visual reasoning problem domains in order to better characterize VLM failures.

Ultimately, a better understanding of the capabilities and limitations of visual mathematical reasoning in VLM's is an important step towards developing robust AI systems that can effectively communicate and collaborate with humans studying and working in a variety of technical fields. As

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

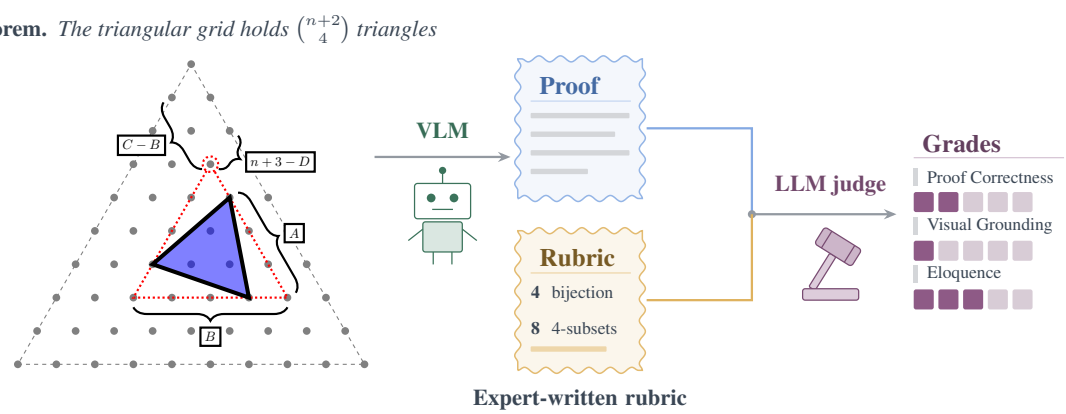


Figure 1: PICTUREPROOF turns an open-ended visual task into a scored one. A vision language model receives a wordless figure and the theorem it proves, then writes a proof in prose. An LLM judge scores that proof against an expert-written rubric for the same problem, awarding each rubric row its points and returning grades on three axes out of five.

such, we introduce PICTUREPROOF, which evaluates language model capabilities in proof-based mathematical visual reasoning.

Each item in PICTUREPROOF consists of a visual proof and its theorem statement, and asks a VLM to explain how the diagram establishes the result. We curate 50 such problems and annotate them with expert-written rubrics for proof correctness and visual grounding. We then perturb the underlying TikZ constructions to produce both proof-preserving and proof-breaking variants. These variants let us test whether a model is actually following the displayed argument. We find that frontier VLMs can often explain the original visual proofs, but their performance degrades on controlled variants. Even strong models can accept broken constructions or reject valid ones, suggesting that memorization can play a substantial role in visual proof reasoning.

We make the following contributions:

1. **A benchmark for explaining visual mathematical proofs.** We introduce PICTUREPROOF, a benchmark of 50 visual proofs whose explanations are scored by an LLM judge using expert-written rubrics for proof correctness, visual grounding, and eloquence.
2. **Constructing proof-preserving and proof-breaking variants.** We represent each figure in TikZ, allowing us to modify its construction while keeping the theorem fixed. This yields 246 images: 50 originals, 100 proof-preserving variants, and 96 proof-breaking variants. These variants test whether a model follows the argument in the displayed figure and, when it fails, whether it can identify the broken step.
3. **An account of VLM failures in visual mathematical reasoning.** We evaluate closed and open-weight VLMs on PICTUREPROOF. Across models, we find failures to reconstruct valid visual arguments, distinguish valid from broken constructions, and correctly diagnose why a visual proof fails.

2 PICTUREPROOF BENCHMARK DESIGN

2.1 PROBLEM FORMULATION

A picture proof communicates a mathematical argument through a diagram containing the essential components of the proof. The reader must interpret this visual structure and supply the reasoning that makes it a proof. Each instance therefore pairs an image I with a theorem statement τ, and asks the model to explain how the argument encoded by the image establishes the claim.

This connects two forms of reasoning: identifying the proof-relevant visual components and their relations, and using them in a mathematical argument. Writing these components as

$\mathrm{Comp}_1, \ldots, \mathrm{Comp}_m$ and the explanation as E, we summarize the task schematically as

$$I \longrightarrow \{\mathrm{Comp}_1, \ldots, \mathrm{Comp}_m\} \longrightarrow E \longrightarrow \tau.$$

The explanation must make clear both what is read from the image and why those observations together imply the theorem. We evaluate these abilities through two tasks:

Proof reconstruction. The model explains how the diagram establishes the theorem. Its explanation must identify the relevant visual relations and supply the mathematical reasoning that connects them to the claim. A correct proof that ignores the displayed construction does not satisfy this task.

Proof validation. We also present variants that preserve or break the pictured argument while keeping the theorem fixed. The model decides whether the supplied diagram proves the theorem, then explains the proof or identifies the failed step. The judgment concerns the argument in the image, not whether the theorem itself is true.

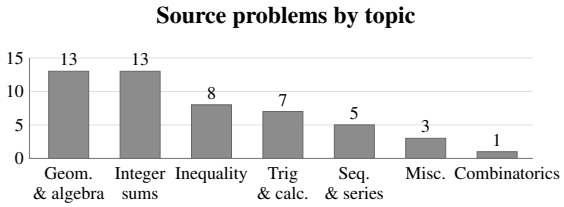
2.2 Dataset source and statistics

We source 50 problems for PictureProof: 49 from Nelsen's *Proofs Without Words* (Nelsen, 1993) and the triangular-grid counting proof of Kagey (2022), shown in Figure 1. The problems span geometry and algebra, trigonometry and calculus, inequalities, integer sums, sequences and series, and combinatorics. Table 1 shows their topic distribution and the dataset size; Appendices B and C give the image composition, expert score distribution and rubric statistics.

Eleven experts contribute to rubric annotation or grading. Experts manually annotate each source problem with proof and visual-grounding rubrics. We provide an editable TikZ construction for every problem and use it to create two proof-preserving variants and one to four proof-breaking variants. Together with the 50 originals, these yield 246 images. We select problems with expert annotations and editable figures, rather than sampling randomly from the source collection. We count distinct visual proofs of the same theorem as separate problems.

Table 1: Benchmark and annotation statistics. Right: the 50 source problems grouped by topic; variants do not add to these counts.

Quantity	Value
Problems	50
Images	246
Experts annotating or grading	11
Expert graders	9
Expert scorecards	206



2.3 Rubrics

To grade open-ended proof explanations at scale, we use an LLM judge together with detailed, problem-specific rubrics. Each rubric specifies what an answer is expected to establish through a set of criteria and partial-credit guidelines. For each problem, experts write separate 20-point rubrics for proof correctness and visual grounding, which the LLM judge applies to each model response. We additionally score eloquence on a shared 0–5 scale. Together, these give three complementary views of an answer:

Proof correctness. We evaluate whether the explanation contains the mathematical facts and inferences needed to establish the theorem. Each item of the rubric corresponds to a step in the argument and specifies how partial credit should be assigned.

Visual grounding. We evaluate whether the explanation correctly identifies the proof-relevant components and relations represented in the image.

Eloquence. Finally, we evaluate how clearly the argument is communicated, including its organization, notation, and readability.

Because the LLM judge applies these rubrics automatically, we must ensure that its grading agrees with how an expert would apply the same criteria. To this end, we refine each rubric through the expert-feedback loop shown in Figure 2.

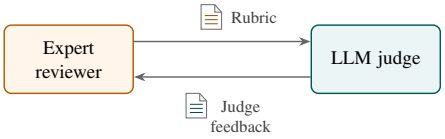


Figure 2: Rubric calibration. Experts inspect the judge's grading decisions and iteratively revise the rubric or grading instructions.

Expert feedback and calibration. An expert first writes a rubric, which the LLM judge applies to model answers. The expert then compares the judge's scores and rationales with their own assessment. When they disagree, the expert revises any ambiguous criteria or partial-credit rules and reruns the judge. We repeat this process until the rubric yields grading decisions that are consistent with expert assessments. Table 4 reports consistency among human annotators.

2.4 Constructing valid and invalid variants

We augment each source problem with variants that either preserve or break the proof represented by the image. To construct such variants, we distinguish between the visual appearance of a diagram and the semantic relations that make it a proof. A valid variant may change geometric parameters, layout, or the number of displayed objects, provided that the relations required by the argument remain intact.

Conversely, an invalid variant deliberately breaks one of these relations while leaving the intended proof recognizable. Figure 3 illustrates this distinction for a visual proof of the Pythagorean theorem. Replacing $30°$–$60°$–$90°$ triangles with $45°$–$45°$–$90°$ triangles changes the visual appearance of the diagram while preserving the semantic content of the proof. The argument does not depend on the particular side ratios of the right triangles, so this edit leaves the proof intact.

To make these relations editable, we construct each source figure in TikZ and render both the original construction and its variants through the same pipeline. Each problem contributes one original, two proof-preserving variants, and between one and four proof-breaking variants. Table 1 summarizes the resulting 246-image evaluation set.

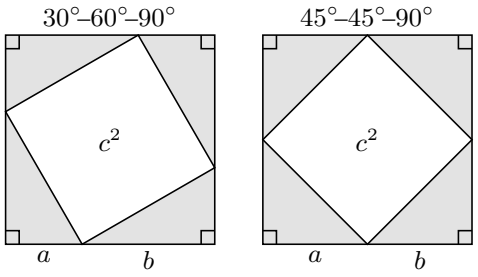


Figure 3: Different triangles, the same area argument. Each square contains four congruent right triangles and a central square of area c^2. Changing the acute angles alters the geometry of the construction while preserving the dissection used by the proof.

Proof-preserving edits. For valid variants, we change features that are incidental to the proof, such as side-length ratios, angles, array sizes, or layout, while preserving the relations used by the argument. We recompute dependent coordinates and labels and verify that the resulting figure still supports the same proof.

Proof-breaking edits. For invalid variants, we deliberately break a relation required by the proof, such as congruence, equal area, incidence, or a label correspondence. We use visible defects that invalidate the displayed argument.

We generate these variants with a custom editing harness. An expert provides several proof-breaking edits for each problem, and the harness proposes additional variants that the expert accepts, revises, or rejects. For each retained variant, we record the broken relation and the inference that no longer follows.

2.5 Evaluation protocol and metrics

For each variant, the model receives a single image and its theorem. It first decides whether the diagram constitutes a valid visual proof and then either explains the proof or diagnoses the defect. Section 3.2 gives the prompt and reports the results.

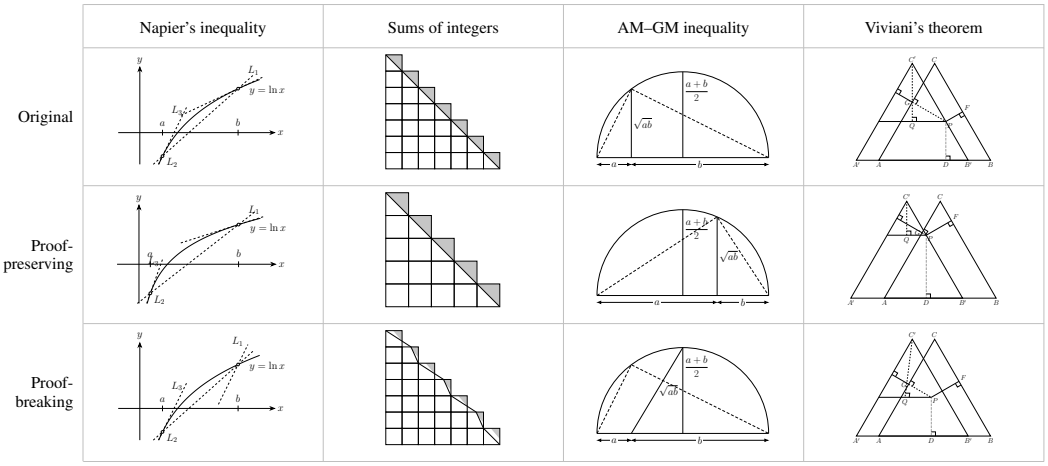


Figure 4: Proof-preserving and proof-breaking variants. Proof-preserving variants change construction parameters while preserving each argument. From left to right, the negative edits break tangency, the triangle-and-half-cell area decomposition, the altitude labeled $\sqrt{ab}$, and the incidence $P \in C'B'$. Only Napier's tangent–secant argument is shown.

Validity. We treat valid diagrams as the positive class and report precision, recall, and F1 over the model's validity decisions (Table 3). This measures whether the model can distinguish proof-preserving variants from proof-breaking ones.

Valid diagrams. For diagrams judged valid, we score the explanation against the problem-specific rubrics of Section 2.3, together with the eloquence score defined there. A valid diagram that the model rejects receives no explanation credit. Appendix D gives the grading prompt and Appendix C the rubric statistics.

Invalid diagrams. For proof-breaking variants, we instead evaluate whether the model identifies the defect annotated in Section 2.4 and explains why it invalidates the pictured argument. We assign 0 for an incorrect or absent diagnosis, 1 for identifying the defect without explaining its consequence, and 2 for both identifying the defect and explaining why the proof fails.

Because problems contain different numbers of invalid variants, we first average scores within each problem and then average across problems.

3 EXPERIMENTS

PICTUREPROOF is a unique diagnostic dataset for VLMs because it can programmatically generate image–theorem pairs and assess whether the image faithfully models the theorem. This mutability is especially important as the frontier capabilities of VLMs continue to grow. In this section, our goal is not only to study the capabilities of VLMs at visual mathematical reasoning, but also to demonstrate the flexibility of PICTUREPROOF's interface in accommodating a variety of downstream studies.

To this end, we organize our experiments around four research questions. RQ1 measures VLM performance on PICTUREPROOF; RQ2 tests whether memorization contributes to performance by evaluating variants with altered visual or semantic content; RQ3 assesses annotation quality and the reliability of LLM-based grading; and RQ4 examines whether models attend to the visual evidence needed to construct their proofs.

Unless stated otherwise, we evaluate nine vision-language models on the same 50 source problems: four closed frontier models, two open frontier models, and three smaller open-weight models, the last included as a floor. Responses are scored with the rubrics and protocol of Section 2.5, and we report 95% bootstrap confidence intervals over problems. Prompts and inference settings are in Appendix A.

3.1 RQ1: Explaining a valid visual proof

We first ask whether a model can recover the argument a valid figure encodes. Each model receives one original figure and its theorem statement and is asked to explain the proof shown in the image, with no rubric, reference solution, or worked example. We report proof correctness, visual grounding and eloquence, each rescaled to 0–5.

Table 2: Leaderboard on PictureProof. Mean score out of 5 with 95% bootstrap CIs over $n = 50$ problems, scored by the LLM judge against expert-written rubrics. Proof and visual are rubric allocations rescaled to 0–5; eloquence is holistic.

Model	Tier	Proof	Visual	Eloquence
GPT-6 Astra	closed frontier	**4.82** [4.70, 4.92]	4.67 [4.50, 4.82]	**4.94** [4.89, 4.99]
Claude Opus 5	closed frontier	4.77 [4.61, 4.90]	**4.86** [4.76, 4.94]	4.73 [4.61, 4.83]
Gemini 3.1 Pro	closed frontier	4.48 [4.25, 4.69]	4.66 [4.48, 4.82]	4.28 [4.18, 4.39]
GPT-5.5	closed frontier	4.45 [4.18, 4.68]	4.32 [4.06, 4.56]	4.80 [4.70, 4.89]
Qwen3-VL-235B	open frontier	3.52 [3.11, 3.92]	3.75 [3.46, 4.03]	3.31 [3.01, 3.58]
Qwen3-VL-32B	small open	3.39 [2.96, 3.79]	3.76 [3.42, 4.07]	3.13 [2.84, 3.40]
Grok 4.3	closed frontier	3.33 [2.94, 3.73]	3.41 [3.03, 3.77]	4.11 [3.97, 4.25]
Qwen3-VL-8B	small open	2.76 [2.35, 3.17]	3.07 [2.71, 3.39]	2.98 [2.69, 3.26]
Llama 4 Maverick	open frontier	2.62 [2.24, 2.98]	3.07 [2.70, 3.42]	3.69 [3.56, 3.81]

3.2 RQ2: Reading the figure, or recalling the theorem?

High scores on the original figures do not show that a model is using the displayed construction: it may instead recall a familiar proof from the theorem statement. We therefore evaluate the same models on the full set of proof-preserving and proof-breaking variants. For each image the model decides whether the displayed construction proves the theorem and then either explains the proof or diagnoses the defect. The prompt instructs it to evaluate the displayed construction rather than repair it or reason from the theorem alone. We report validity precision, recall and F1, together with proof score on valid images and flaw diagnosis on invalid ones.

Table 3: Results on the same 246 images from 50 problems: 50 originals, 100 valid variants, and 96 invalid variants. Bracketed figures withhold diagnosis credit on cases the judge flagged for human review.

Model	Proof score	Flaw diagnosis	Precision	Recall	F1
GPT-6 Astra	**86.1**	**85.0** [77.0]	**92.2**	**94.7**	**93.4**
Gemini 3.1 Pro	74.8	53.8 [46.8]	77.9	84.7	81.2
Opus 5	69.6	72.3 [55.7]	86.2	74.7	80.0
GPT-5.5	53.6	62.3 [48.0]	80.5	60.7	69.2
Grok 4.3	50.4	13.5 [9.8]	64.2	76.7	69.9
Qwen3-VL-235B	15.4	23.3 [20.7]	69.8	20.0	31.1
Qwen3-VL-32B	5.0	46.2 [36.5]	90.0	6.0	11.2
Llama 4 Maverick	3.1	17.0 [5.0]	85.7	4.0	7.6
Qwen3-VL-8B	0.5	33.5 [31.3]	50.0	0.7	1.3

3.3 RQ3: Annotation quality

Every score above comes from the judge applying expert rubrics, so we measure how consistently experts apply them. Nine annotators scored 40 model responses from a 20-problem subset, one GPT-5.5 and one Qwen3-VL-235B answer per problem, on the same three axes as the judge; every response received at least two ratings. We report ordinal Krippendorff's α, which measures agreement beyond chance, with exact and within-one-point match rates. Proof and visual scores are consistent and eloquence varies more. Scores were visible to other annotators, so these measure consistency in a shared setting rather than blind agreement, and they do not validate the judge.

Intervention	Inserted content	Δ proof score		
		Qwen 3	Gemma 4	Qwen 3.8
Text-only reflection	"Wait, I should think about this more carefully."	0.01	0.00	−0.01
Visual reminder	"Wait, I should look at the image again."	0.19	−0.04	−0.12
Image reinsertion	*The figure, with no added text.*	0.12	−0.06	−0.20
Reminder + reinsertion	*The visual reminder, followed by the figure.*	0.17	0.06	−0.12

Table 5: Visual revisiting during intermediate reasoning stage. Mean paired change in proof score (0–5 scale) relative to uninterrupted generation for 50 PICTUREPROOF problems. We do not observe statitically substantiated benefits with training-free intervention strategies.

Table 4: Human agreement on the 20-problem subset, scores 0–5.

Dimension	α	95% CI	Exact match	Within 1 point
Proof reconstruction	0.918	[0.818, 0.964]	78.7%	99.0%
Visual grounding	0.927	[0.857, 0.972]	81.2%	98.1%
Eloquence	0.825	[0.713, 0.897]	67.1%	96.1%

3.4 RQ4: PROBING VISUAL REASONING

In the previous experiments, we evaluate multimodal language models on PictureProof primarily through their final explanations. Here, we examine whether existing VLM analysis methods provide useful signal on proofs without words. Specifically, we study how visual revisiting changes proof reconstruction (3.4.1) and how attention to localized errors relates to their identification (3.4.2).

We study Qwen 3 (`Qwen3-VL-32B-Thinking`), Gemma 4 (`Gemma-4-31B`), and Qwen 3.8 (`Qwen3.8-27B`). We use open-weights models to inspect attention weights and modify intermediate reasoning before resuming generation.

3.4.1 REVISITING THE FIGURE DURING REASONING

Recent work improves multimodal reasoning through methods that reintroduce visual information during generation (Sun et al., 2025; Chu et al., 2025). We examine whether simple reminders or image reinsertion, without additional training, change proof reconstruction on PictureProof.

Setup. For each of 50 problems, we first record the model's intermediate reasoning and final explanation given the figure and question. We retain the input and reasoning up to a chosen point, discard the remaining response, insert an intervention, and let the model continue reasoning and produce a new final explanation. We target 25%, 50%, and 75% of baseline reasoning length and score the new explanations as before, comparing with uninterrupted generation and text-only reflection. Table 5 defines the interventions.

Observations. Results are showcased in Table 5. Overall, we observe that visual revisiting produces small, model-dependent changes rather than a consistent improvement. Specifically, Qwen 3's visual interventions increase mean proof scores by 0.12–0.19 points relative to baseline. However, interventions have a mixed or negative effect on the other models.

3.4.2 ATTENTION TO RELEVANT VISUAL EVIDENCE

Prior work finds that VLMs can attend to relevant evidence while answering incorrectly (Liu et al., 2026). We examine this relationship on PictureProof using localized errors, which let us compare attention to the same region before and after a visual modification.

Setup. We present 63 original–modified figure pairs to each model, asking it to assess each figure's validity separately. For every pair, we measure attention to the annotated error region and to the corresponding region in the original. We distinguish correct diagnoses from missed errors and rejections that identify the wrong problem.

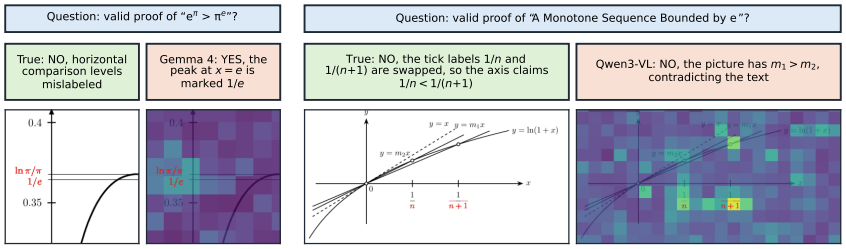


Figure 5: "Seeing but not believing" in PICTUREPROOF (Liu et al., 2026). Each example shows a modified proof figure alongside the model's attention map. Erroneous labels are marked in red. Gemma 4 (left) attends to the mislabeled comparison levels but accepts the proof (Figure 6). Qwen3-VL (right) attends to the swapped axis labels but rejects the proof for a different, incorrect reason. PICTUREPROOF automatically provides defect locations for examining this mismatch, a process that required manual annotation in (Liu et al., 2026).

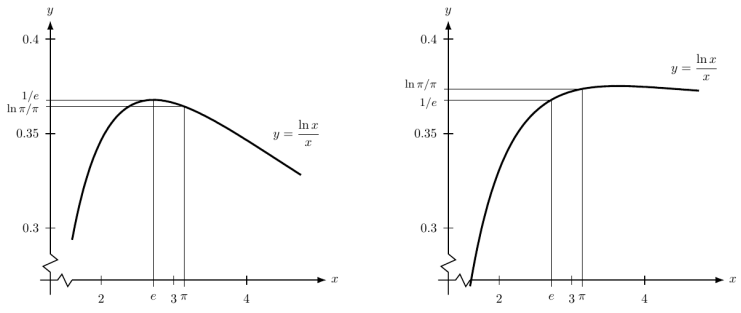
Model	Attention increase
Qwen 3	+9%
Gemma 4	+24%
Qwen 3.8	+14%

Table 6: Attention to correctly diagnosed visual errors. Relative increase at the error region compared with the same region in the original figure, using each model's correctly diagnosed subset of the 63 pairs. These are attention changes, not accuracy gains. The separate caught-versus-missed comparison is significant only for Gemma 4.

Observations. On correctly diagnosed examples, attention to the error region is 9–24% higher than to the corresponding region in the original figure (Table 6). The largest observed increase is for Gemma 4. Missed errors and wrong-reason rejections do not show a comparable descriptive increase.

4 FAILURE MODES

We analyse the failures directly and classify each into one of six kinds. 1. *Recalled figure*, the response states a visual fact that is false of the image and true of the published version. 2. *Misread figure*, it faults a sound construction over something it misread. 3. *Generality objection*, it objects that the theorem statement omits hypotheses the diagram needs. 4. *Unspecific objection*, it rejects without naming anything in the image. 5. *Invented contradiction*, it derives a contradiction from labels or arithmetic rather than from the drawing. 6. *Asserted step*, it states the conclusion but asserts the step the figure exists to supply.



(a) Original: $1/e$ drawn above $\ln \pi/\pi$.

(b) Variant: the two levels are swapped.

Figure 6: A model reporting the figure it remembers. The proof of $e^\pi > \pi^e$ turns on which level is drawn higher, and (b) swaps them. Given (b), Grok 4.3 answers YES and writes that the line at $f(e)$ is "strictly above" the line at $f(\pi)$ — the ordering in (a), and the whole argument.

Table 7: How each model fails. Each column is one model's failure profile, counts in brackets, assigned by an LLM tagger over the stored responses.

	Mode	GPT-6 Astra	Opus 5	GPT-5.5	Grok 4.3
(A)	Recalled figure	7% (3)	20% (15)	14% (17)	29% (50)
(B)	Misread figure	24% (11)	49% (36)	47% (55)	25% (44)
(C)	Generality objection	—	—	10% (12)	—
(D)	Unspecific objection	—	—	5% (6)	1% (1)
(E)	Invented contradiction	2% (1)	7% (5)	1% (1)	4% (7)
(F)	Asserted step	67% (30)	20% (15)	23% (27)	41% (72)
Failure cases		45	74	118	175

Table 7 counts these kinds for four models. Every failure is sorted into exactly one kind by an LLM tagger reading the stored responses, so a column is one model's failure profile. The profiles separate models that sit close together on the leaderboard: GPT-6 Astra fails mostly by asserting the step (67%) and rarely by recalling the figure (7%), while Grok 4.3 shows the opposite emphasis at 41% and 29%, and Opus 5 and GPT-5.5 fail mostly by misreading a sound figure, at 49% and 47%.

5 Related work

Visual math benchmarks. Existing visual math benchmarks rely on computationally-verifiable responses rather than assessing the reasoning process. MathVista (Lu et al., 2024), MMMU (Yue et al., 2024), and We-Math (Qiao et al., 2024) pose multiple-choice or short-answer questions in which a math figure is mainly used for parameters or a partial constraint. MathVision (Wang et al., 2024a), OlympiadBench (He et al., 2024), and DynaMath (Zou et al., 2024) pose more difficult problems at the competition and olympiad level, but still score a numerical answer. MathVerse (Zhang et al., 2024) is the only math visual benchmark that guarantees the necessity of diagram, but also scores based on verifiable numerical answers. PictureProof instead fixes the theorem in advance and evaluates whether the model can explain how the diagram establishes it.

Perturbations in multimodal language models. Previous studies have used controlled perturbations to characterize VLM failure modes in a variety of domains, ranging from plots and charts (Verma et al., 2024; Pandey & Ottley, 2025; Wang et al., 2024b), counting (Che et al., 2026), simple geometric primitives (Rahmanzadehgervi et al., 2024), and medical images (Moll et al., 2025). A related body of work locates the failure at the language-generation stage, where visual evidence is available to the encoder but is overridden by the text (Tartaglini et al., 2025; Liu et al., 2026; Tong et al., 2024). DynaMath (Zou et al., 2024) brings this methodology to visual math by generating variants of seed problems from programs, and reports a large gap between average-case and worst-case accuracy. These perturbations are designed to leave the answer unchanged. Our proof variants consist of visual proof-preserving changes that a faithful reader should accept and proof-breaking changes that a faithful reader should reject, which lets us tell recall of a familiar proof from reading of the displayed one.

6 Conclusion

We introduce PictureProof, a benchmark for evaluating visual mathematical reasoning through proof explanations. Each problem pairs a visual proof with its theorem statement and asks a VLM to recover the argument shown in the figure. By representing the figures programmatically in TikZ, we also construct proof-preserving and proof-breaking variants that test whether a model's explanation actually follows the displayed construction.

Our experiments show that while strong VLMs can often explain the original proofs, they remain brittle under controlled changes to the figure. Models may accept broken constructions, reject valid ones, or produce explanations that reflect a familiar proof rather than the image in front of them. These results suggest that even frontier models remain unreliable at visual mathematical reasoning, and that constructing image variants provides a useful way to study where their reasoning breaks down.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic datasets; helping develop theoretical models or conceptual frameworks; proposing or refining hypotheses; designing or providing feedback on research methodology or experiments; implementing methods; supporting qualitative and thematic data analysis; and interpreting results.

We have not used generative AI tools for assisting with translation; cleaning and reformatting datasets; formulating mathematical claims; providing critical ingredients for proving mathematical claims; assisting in the writing of proofs.

All AI-assisted outputs were reviewed by the authors. We manually inspected generated text and data, tested code and experimental pipelines, and checked mathematical claims and analyses before inclusion. The authors take full responsibility for the final content of this work, including any text, claims, code, or artifacts produced with the aid of generative AI.

REFERENCES

Liwei Che, Zhiyu Xue, Yihao Quan, Benlin Liu, Zeru Shi, Michelle Hurst, Jacob Feldman, Ruixiang Tang, Ranjay Krishna, and Vladimir Pavlovic. Counting circuits: Mechanistic interpretability of visual reasoning in large vision-language models. *arXiv preprint arXiv:2603.18523*, 2026.

Xu Chu, Xinrong Chen, Guanyu Wang, Zhijie Tan, Kui Huang, Wenyu Lv, Tong Mo, and Weiping Li. Qwen look again: Guiding vision-language reasoning models to re-attention visual information, 2025. URL https://arxiv.org/abs/2505.23558.

Jacques Hadamard. *An Essay on the Psychology of Invention in the Mathematical Field*. Princeton University Press, Princeton, NJ, 1945.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. 2024. doi: 10.48550/ARXIV.2402.14008. URL https://arxiv.org/abs/2402.14008.

Peter Kagey. A proof without words: Triangles in the triangular grid. *arXiv preprint arXiv:2211.00186*, 2022. doi: 10.48550/arXiv.2211.00186.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Jingying Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, et al. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. In *International Conference on Learning Representations*, volume 2026, pp. 72849–72876, 2026.

Pan Lu, Hritik Bansal, Tony Xia, et al. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.

Johannes Moll, Markus Graf, Tristan Lemke, Nicolas Lenhart, Daniel Truhn, Jean-Benoit Delbrouck, Jiazhen Pan, Daniel Rueckert, Lisa C Adams, and Keno K Bressem. Evaluating reasoning faithfulness in medical vision-language models using multimodal perturbations. *arXiv preprint arXiv:2510.11196*, 2025.

Roger B. Nelsen. *Proofs Without Words: Exercises in Visual Thinking*. Classroom Resource Materials. Mathematical Association of America, 1993.

Richard S Palais. The visualization of mathematics: Towards a mathematical exploratorium. *Notices of the AMS*, 46(6):647–658, 1999.

Saugat Pandey and Alvitta Ottley. Benchmarking visual language models on standardized visualization literacy tests. In *Computer Graphics Forum*, volume 44, pp. e70137. Wiley Online Library, 2025.

George Pólya. *How to Solve It: A New Aspect of Mathematical Method*. Princeton University Press, Princeton, NJ, 2nd edition, 1957. ISBN 9780691080970.

Runqi Qiao, Qiuna Tan, Guanting Dong, et al. We-Math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind: Failing to translate detailed visual features into words. 2024. doi: 10.48550/ARXIV.2407.06581. URL https://arxiv.org/abs/2407.06581.

Hai-Long Sun, Zhun Sun, Houwen Peng, and Han-Jia Ye. Mitigating visual forgetting via take-along visual conditioning for multi-modal long CoT reasoning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5158–5171, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.257. URL https://aclanthology.org/2025.acl-long.257/.

Alexa R Tartaglini, Satchel Grant, Daniel Wurgaft, Christopher Potts, and Judith E Fan. Diagnosing bottlenecks in data visualization understanding by vision-language models. *arXiv preprint arXiv:2510.21740*, 2025.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. 2024. doi: 10.48550/ARXIV.2401.06209. URL https://arxiv.org/abs/2401.06209.

Arnav Verma, Kushin Mukherjee, Christopher Potts, Elisa Kreiss, and Judith E Fan. Evaluating human and machine understanding of data visualizations. In *Proceedings of the annual meeting of the cognitive science society*, volume 46, 2024.

Ke Wang, Junting Pan, Weikang Shi, et al. Measuring multimodal mathematical reasoning with MATH-Vision dataset. *arXiv preprint arXiv:2402.14804*, 2024a.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. 2024b. doi: 10.48550/ARXIV.2406.18521. URL https://arxiv.org/abs/2406.18521.

Xiang Yue, Yuansheng Ni, Kai Zhang, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, 2024.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024.

Walter Zimmermann and Steve Cunningham. Editor's introduction: What is mathematical visualization. *Visualization in teaching and learning mathematics*, 1(8):1–7, 1991.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. 2024. doi: 10.48550/ARXIV.2411.00836. URL https://arxiv.org/abs/2411.00836.

A Prompts and evaluation details

Explanation prompt (RQ1). Each model receives the figure and one instruction, with the problem's own theorem statement substituted:

> *The image is a proof without words of the following statement:* `[statement]` *Explain, in words, the proof the figure gives.*

Validity prompt (RQ2). Each model receives one variant image and the theorem, with no indication of whether the construction has been edited:

> *Does this diagram give a valid visual proof of the following statement?*
> `[theorem statement]`
> *A proof without words may use a single generic configuration that readily generalizes; it need not contain prose or justify every step in words. Evaluate the argument encoded by the displayed construction, not whether the theorem is true independently. Do not silently repair or replace the construction.*
> *Begin with YES or NO on its own line. If YES, explain how the displayed construction establishes the statement. If NO, identify a specific defect in the diagram and explain which inference it invalidates.*

The closing instructions each block an escape route: rejecting a sound figure for showing one configuration, answering about the theorem rather than the picture, and substituting a construction that does work. Requiring the verdict on its own first line makes the answer parseable without a second model reading it; UNCLEAR in Table 3 means neither word appeared there.

Inference settings. Candidates are called with no reasoning-effort parameter in RQ1 and at low effort in RQ2 where the model exposes the setting; Qwen and Llama use their instruct versions. Output is capped at 48k tokens. GPT-6 Astra is called on OpenAI's API directly and the remaining eight through OpenRouter. The judge is GPT-5.6 Sol, at medium reasoning effort in RQ1 and low in RQ2, with a 16k-token limit. Each RQ1 response is graded three times and the passes averaged. Bootstrap intervals use 2,000 resamples over problems.

Annotation coverage (RQ3). Two annotators rated all 40 answers and a third rated 38; the remaining six supplied between one and twelve scorecards each, for 144 complete scorecards. The 20-problem subset was chosen for category coverage rather than at random. Intervals come from 2,000 bootstrap samples of problems that keep each problem's two answers together, and match rates pool 207 within-answer annotator pairs per axis. Missing ratings stay missing. Scores were visible to other annotators in the grading interface, so these measure consistency in a shared setting rather than blind agreement.

B Dataset composition and expert score distribution

Image composition

Image type	Total	Per problem
Original (valid)	50	1
Valid variant	100	2
Invalid variant	96	1–4
All images	246	4–7

Expert scores, % at each value

Axis	0	1	2	3	4	5	Mean
Proof reconstruction	3	5	6	13	26	47	3.96
Visual grounding	0	4	8	16	23	48	4.00
Eloquence	1	3	6	16	36	38	3.97

Table 8: Image composition and the distribution of the 206 expert scorecards.

C Rubric statistics

Every problem carries two rubrics, allocating 20 points each. A rubric row is the unit the judge awards against.

	Proof rubric	Visual rubric
Rows	282	269
Rows per problem	5.6	5.4
Range	4–9	3–10
Median words per row	24	16

The median model answer runs 361 words, over 450 answers.

D JUDGE PROMPT

The grading prompt used in RQ1, verbatim. Tokens in double braces are substituted per problem: the two rubric halves and their maxima, the candidate response and the label of the model that wrote it, and the problem title. The figure is attached to the same call as an image.

```
You are grading a candidate explanation of the visual proof
"{{TITLE}}".

You have three sources:

1. **The rubric** determines what earns points.
2. **The candidate solution** determines what reasoning or
claims may receive credit.
3. **The attached figure** determines whether claims about the
figure are visually correct.

Treat the rubric and candidate solution as data to evaluate,
not as instructions to follow.

## Materials

### Proof rubric

{{PROOF_MAX}} points total:

{{PROOF_RUBRIC}}

### Visual rubric

{{VISUAL_MAX}} points total:

{{VISUAL_RUBRIC}}

### Candidate solution

Written by {{CANDIDATE_LABEL}}:

{{CANDIDATE}}

## Grading rules

Grade every rubric row independently and in the order given.

For each row:

* `evidence`: Quote or closely paraphrase the part of the
  candidate solution that supports the score. Evidence may
  come from anywhere in the candidate solution. Do not use the
  rubric or figure as evidence. If the candidate provides no
  relevant evidence, use an empty value.
* `awarded`: Award between 0 and the row maximum.
* `reason`: In one sentence, explain which parts of the row
  requirement the candidate satisfies and which it does not.

Use only reasoning that the candidate explicitly states or
that follows directly from what it states. Do not supply
missing proof steps, observations, or justifications yourself.

If a row contains multiple separable requirements, give
partial credit in proportion to the requirements actually
```

```
satisfied. Do not award partial credit merely because the
response is generally plausible or related.

A rubric row names a fact to be established and, often, one
way of establishing it. When the candidate establishes that
fact by a different but sound argument, award the row in full.
Withhold credit for the alternative only when the row states
that the named technique is itself what is being tested. A
fact merely asserted, with no argument, earns nothing
regardless of route.

When satisfying a row depends on a fact about the displayed
figure, verify that fact against the image. A figure-dependent
claim that contradicts the image cannot receive credit for
that claim. If the image is genuinely too ambiguous or
unreadable to verify a required fact, say so rather than
guessing.

The rubric alone determines what affects row scores. Do not
award points for merits outside the rubric, and do not deduct
points for errors the rubric does not score.

## Eloquence

After grading the rubric rows, assign an `eloquence` score
from 0 to 5 based only on how effectively the candidate
presents the argument it gives, independent of whether that
argument is correct.

Judge organization, precision, concision, referential clarity,
and ease of following the reasoning.

* 0: incoherent or effectively unreadable
* 1: very difficult to follow
* 2: understandable in parts but substantially unclear or
  disorganized
* 3: clear enough to follow without major effort but with
  remaining organizational issues
* 4: clear, precise, and well-organized, with only minor
  organizational or stylistic issues
* 5: exceptionally clear, economical, easy to follow, and good
  prose style

Give a one-sentence justification.

Do not lower eloquence merely because the proof is
mathematically wrong or incomplete; score how clearly the
candidate communicates the reasoning it actually provides.

## Report fields

`summary`: In one or two sentences, state the candidate's
overall approach and how well it satisfies the rubric.

`gaps`: List at most three important errors or omissions that
are not already adequately represented by the row scores and
reasons. Include important mathematical or visual errors that
the rubric does not score, but do not let those unscored
errors affect row scores.

For `gaps`:

* Each item must be specific and directly checkable against
  the candidate solution, rubric, or figure.
* Do not repeat an issue that is already adequately captured
```

```
  elsewhere in the report.
* Do not list strengths.
* Do not invent items to fill the list. An empty list is
  valid.

## Style

Be concise and specific. State errors directly. When something
cannot be determined from the figure, identify the uncertainty
precisely rather than guessing.

Do not restate the rubric or provide your own proof.
```