GEOPERTURB: PROBING STRUCTURAL VULNERABILITY IN MULTIMODAL GEOMETRY REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Multimodal Models achieve strong results on visual geometry benchmarks, yet it remains unclear whether this performance reflects geometric reasoning that remains stable across equivalent drawings. Rather than relying on low-level pixel corruptions, we introduce GEOPERTURB, a matched diagnostic framework that reconstructs each problem in TikZ and generates four answer-preserving variants that change orientation, symbol identity, visual appearance, or irrelevant structure. The resulting benchmark contains 531 matched problem suites. We combine paired accuracy changes with canonical-conditioned retention to identify solved-to-failed transitions that marginal accuracy can hide. Across proprietary and open-source LMMs, we find that the same problem is often answered correctly in one equivalent view and incorrectly in another. This instability is selective rather than generic: rotation and relabeling expose the clearest failures, while appearance changes and irrelevant clutter are less disruptive for frontier models. A controlled fine-tuning study further shows that explicit derivation supervision, rather than perturbation exposure alone, drives cross-view improvement. GEOPERTURB therefore shifts geometry evaluation from asking whether a model can answer one drawing to asking whether it preserves the same solution across equivalent drawings. The central weakness is not simply low accuracy, but unstable correctness.Code is available at `https://anonymous.4open.science/r/geoperturb_minimal-F855`; datasets will be released publicly.

1 INTRODUCTION

The rapid convergence of computer vision and natural language processing has fueled the rise of Large Multimodal Models (LMMs), positioning them as core engines for advanced intellectual tasks (Yan et al., 2025; Zhong et al., 2025). Among these, automated mathematical reasoning, specifically geometric problem solving, serves as a crucial frontier for evaluating an agent's high-order cognitive capabilities (Xu et al., 2024; Chen et al., 2025; Li et al., 2026; Cho et al., 2026; Ping et al., 2026). Geometric deduction requires a sophisticated interplay between perceptual parsing and logical execution: a model must not only recognize visual primitives such as lines, angles, and curves from a diagram but also map them precisely to the abstract mathematical constraints articulated in the textual prompt. Driven by massive-scale pre-training and alignment, current state-of-the-art LMMs (Zhan et al., 2025; Yang et al., 2026a) have achieved unprecedented scores on static geometry benchmarks, sparking optimism that these systems have begun to master genuine mathematical deduction (Yang et al., 2026b). Yet these scores answer only whether a model can solve one drawing. Geometry requires a stronger property: once a problem is solved, the solution should survive any redraw that preserves the same geometric relations and answer. Single-view accuracy does not test this property.

Prior work shows that language models can rely on superficial patterns instead of consistent deduction (Wang et al., 2021; Zhu et al., 2024). Many multimodal robustness tests focus on pixel corruption or image degradation (Cui et al., 2024; Qiu et al., 2022). Such tests measure resistance to visual noise, but geometry demands a different form of robustness. A diagram is one rendering of an underlying mathematical structure: when the structure and constraints remain unchanged, the solution should remain unchanged. Our question is therefore simple: *when the drawing changes but the geometry does not, does the model preserve its answer?*

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.

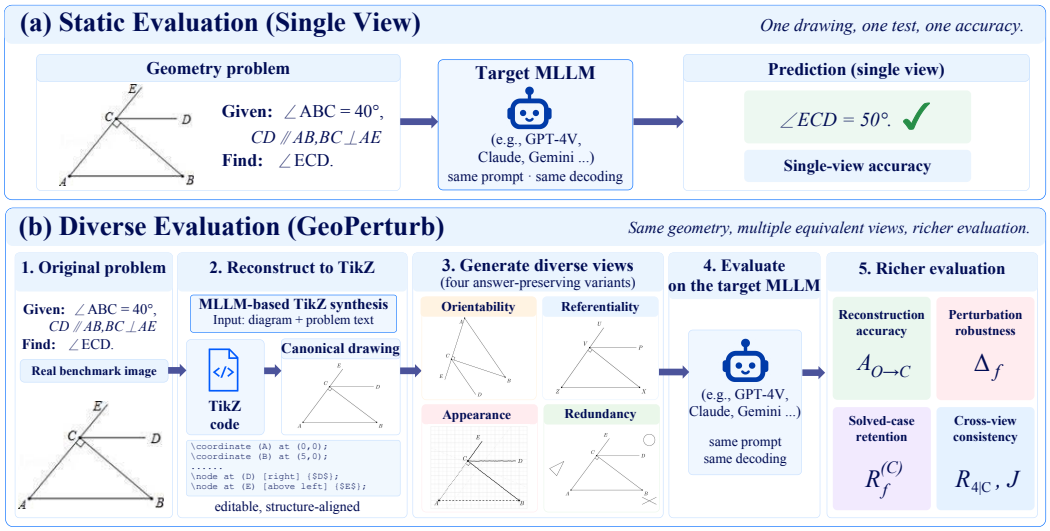

Figure 1: The GEOPERTURB framework: TikZ reconstruction enables four answer-preserving interventions and matched evaluation of whether solved problems remain solved.

To bridge this gap, we present GEOPERTURB, a programmatic framework that changes the unit of evaluation from one fixed drawing to a matched suite of equivalent views. Instead of treating diagrams as immutable pixel arrays, we reconstruct each source diagram into editable TikZ and change one structural factor at a time while keeping the underlying problem and intended solution fixed. We consider four answer-preserving interventions: Orientability (spatial rotation), Referentiality (symbolic permutation), Appearance (visual stylization), and Redundancy (irrelevant injection). This same-problem design turns redrawing into a controlled intervention: a correct-to-incorrect transition directly reveals a failure to preserve the solution across equivalent representations.

Our empirical evaluation reveals a clear vision–logic gap: for current LMMs, a geometry problem is not simply solved or unsolved; the same problem can be solved under one equivalent drawing and missed under another. This instability is structured rather than generic: rotation and relabeling expose the clearest failures, while appearance changes and irrelevant clutter are substantially less disruptive for frontier models. Moreover, marginal accuracy can hide these failures because gains on previously incorrect problems can offset losses on problems already solved; canonical-conditioned retention exposes these solved-to-failed transitions. A controlled fine-tuning study further shows that derivation-supervised finetuning on perturbed views improves cross-view performance, whereas answer-only finetuning on the same views does not.

In summary, our main contributions are:

- **Introducing Matched Invariance Evaluation:** We introduce a geometry-specific evaluation protocol that tests the same problem across answer-preserving views, shifting evaluation from one fixed drawing to whether a solved problem remains solved under controlled redraws.
- **Constructing GeoPerturb:** We construct 531 audited problem suites, each containing a Canonical TikZ reconstruction and four controlled variants that change orientation, symbol identity, visual appearance, or irrelevant structure.
- **Revealing Structured Instability:** We combine paired family effects with canonical-conditioned retention to expose solved-to-failed transitions that can remain hidden in marginal accuracy. Across models, rotation and relabeling expose the clearest failures, while appearance changes and irrelevant clutter are less disruptive for frontier models.

2 RELATED WORK

Reasoning Benchmarking. Recent advancements in LMMs have significantly expanded the horizon of automated reasoning, particularly in complex domains such as geometry (Xu et al., 2024; Zhao et al., 2025; Fu et al., 2025; Shi et al., 2025; Zhang et al., 2025; Wu & Sun, 2025). Existing

benchmarks including GeoQA (Chen et al., 2021) and Geometry3K (Lu et al., 2021) have been instrumental in assessing the ability of models to translate visual information into mathematical proofs. However, these benchmarks predominantly evaluate each problem under a fixed diagram, so strong performance does not reveal whether the same solution is preserved under equivalent redraws. This leaves an important aspect of geometric reasoning untested: invariance to representation changes that preserve the underlying problem and answer. GEOPERTURB complements existing benchmarks by directly evaluating this within-problem invariance.

Robustness Probing. The robustness of vision-language models has been studied through adversarial attacks (Goodfellow et al., 2014; Qi et al., 2024), image corruptions (Hendrycks & Dietterich, 2019), and cross-modal reasoning conflicts (Zhu et al., 2026). IsoBench (Fu et al., 2024) compares visual, textual, and mathematical representations of the same problems. Our focus is whether the same geometric solution survives an equivalent redraw within the image–text setting. Most closely related, DynaMath (Zou et al., 2025) generates visual-math variants by changing problem conditions and visual appearance. The key difference is the controlled comparison: DynaMath evaluates robustness across variants of a seed question, whereas GEOPERTURB fixes the geometric conditions and intended answer to test whether a solved problem remains solved after redrawing. Our contribution is therefore a matched within-problem invariance test, rather than dynamic instance generation itself.

Programmatic Synthesis. The integration of programmatic data generation, particularly using languages like TikZ, has emerged as a powerful paradigm for augmenting training data and enhancing compositional reasoning (Belouadi et al., 2024a;b; Wei et al., 2025; Greisinger & Eger, 2026). By leveraging code as a medium to define geometric properties, researchers have successfully generated synthetic datasets that bypass the limitations of human-annotated samples (Saito et al., 2025; Sun et al., 2026). While these methods primarily use programmatic representations for generation and training, their use as controlled diagnostic interventions remains less explored. GEOPERTURB uses programmatic reconstruction differently: the editable TikZ representation enables matched, answer-preserving interventions that vary one aspect of the diagram while keeping the underlying problem fixed.

3 THE GEOPERTURB FRAMEWORK

3.1 PROBLEM FORMULATION

Let problem i be $P_i = (I_i^{\text{orig}}, T_i, A_i)$, where I_i^{orig} is the original diagram, T_i is the textual problem statement, and A_i is the ground-truth answer. We reconstruct the original diagram into a canonical TikZ view I_i^{can} and apply perturbation family $f \in \mathcal{F}$ to obtain a matched variant $(I_{i,f}, T_{i,f}, A_{i,f})$. For Orientability, Appearance, and Redundancy, $T_{i,f} = T_i$ and $A_{i,f} = A_i$. Referentiality applies a synchronized bijection to diagram labels and any label-dependent text or answer representation while preserving the underlying geometric problem and intended solution.

For a fixed model $\mathcal{M}$, we define correctness indicators

$$O_i = \mathbf{1}\Big[\mathcal{M}(I_i^{\text{orig}}, T_i) = A_i\Big],$$
$$C_i = \mathbf{1}[\mathcal{M}(I_i^{\text{can}}, T_i) = A_i],$$
$$F_{i,f} = \mathbf{1}[\mathcal{M}(I_{i,f}, T_{i,f}) = A_{i,f}].$$

For perturbation family f, let $\mathcal{V}_f$ denote the set of valid Canonical–variant pairs and $n_f = |\mathcal{V}_f|$. We measure the family-level effect using the matched accuracy change

$$\Delta_f = \frac{1}{n_f} \sum_{i \in \mathcal{V}_f} (F_{i,f} - C_i).$$

Thus, Δ_f compares the perturbed and Canonical views on exactly the same problem IDs.

To measure whether a problem that is already solved remains solved after the intervention, we report canonical-conditioned retention

$$R_f^{(C)} = \widehat{\Pr}(F_f = 1 \mid C = 1) = \frac{\sum_{i \in \mathcal{V}_f} C_i F_{i,f}}{\sum_{i \in \mathcal{V}_f} C_i}.$$

An ideally invariant model has $R_f^{(C)} = 1$ and $\Delta_f = 0$ for every valid perturbation family. We therefore characterize structural stability using these matched family effects and conditional retention, with problem-level confidence intervals reported for the corresponding estimates.

3.2 Perturbation Taxonomies

We construct four families of answer-preserving transformations, illustrated in Figure 1. Each family changes one aspect of the representation while preserving the underlying geometric problem and intended solution. We first ran family-specific checks on each variant. Three independent annotators then reviewed the rendered views. We corrected flagged cases and checked them again before release.

- **Orientability** (*Spatial Rotation*): We apply discrete rigid rotations to the geometric structure while keeping vertex labels upright, legible, and attached to the same entities. This tests orientation sensitivity while keeping vertex names upright and legible.
- **Referentiality** (*Symbolic Permutation*): We apply a one-to-one permutation of vertex labels and synchronize the mapping across the diagram, question, options, and label-dependent answer representation when required. The underlying geometry and intended solution remain unchanged, testing whether symbol-to-entity grounding remains stable when the names of geometric entities change.
- **Appearance** (*Visual Stylization*): We vary grids, colors, line styles, and sketch rendering while preserving answer-critical topology and labeled relations. Variants are rejected if a style change obscures a required primitive or introduces a proof-relevant mark or constraint.
- **Redundancy** (*Irrelevant Injection*): We add detached, unlabeled auxiliary primitives in vacant canvas regions. Retained objects have no contact, intersection, tangency, collinearity, labeling, or proof-relevant alignment with the core construction, so this family tests selectivity under irrelevant visual structure without adding new geometric constraints.

3.3 Data Construction

We construct 531 matched problem suites from GeoQA (Chen et al., 2021), Geometry3K (Lu et al., 2021), and an independently collected Pure set. Each suite contains the original source raster, a canonical TikZ reconstruction, and one audited variant from each of the four perturbation families. We refer to the original benchmark partitions as **Source-Train** ($N = 347$) and **Source-Test** ($N = 96$) to distinguish them from training and testing of the evaluated LMMs. The **Pure** split ($N = 88$) was independently collected from public advanced-geometry examination material, OCR-processed, cleaned, normalized, and converted into our data format; it was not sampled from GeoQA, Geometry3K, MathVista, or other mainstream multimodal benchmarks. We use Pure as an independently constructed out-of-distribution evaluation split.

We independently annotated all 2,655 views: 531 Canonical views and 2,124 perturbation variants. Three graduate students in mathematics education reviewed each view before adjudication. We corrected every non-VALID case and re-audited it before fixing the release. The pre-adjudication agreement was Fleiss' $\kappa = 0.87$, with 98.6% raw agreement; the full protocol is given in Appendix A.

4 Experiments

Evaluated Models. To ensure a comprehensive and representative diagnosis of contemporary vision-language capabilities, we construct an evaluation suite spanning both industry-leading proprietary foundations and state-of-the-art open-source architectures. Our evaluation matrix includes:

- **Proprietary Models:** We benchmark a diverse cohort of commercial closed-source models, evaluating the frontier capabilities of `GPT-5.4-High` (OpenAI, 2026), `GPT-4o-mini` (OpenAI, 2024), `Claude-4.6-Opus` (Anthropic, 2026a), `Claude-4.6-Sonnet` (Anthropic, 2026b), `Gemini-3.1-Pro-Preview` (Google DeepMind, 2026), and `Qwen3-VL-Plus` (Bai et al., 2025).
- **Open-Source Models:** To facilitate reproducibility and analyze accessible structural representations, we extend our evaluation to prominent open-source vision-language models across varying parameter scales, including the `Qwen3-VL-8B` (Bai et al., 2025), the `InternVL3.5-8B` (Wang et al., 2025b), and the unified multimodal baseline `LLaVA-OneVision-1.5-8B` (An et al., 2025).

Table 1: **Macro Benchmarking Results.** Original and Canonical accuracy (%) by source partition. The Original-to-Canonical comparison is treated as a separate representation intervention, while Canonical-to-perturbation effects are analyzed at the micro level. Source-Train and Source-Test refer to the original benchmark partitions, not training and testing of the evaluated models.

Model	Original			Canonical		
	Source-Train	Source-Test	Pure	Source-Train	Source-Test	Pure
Proprietary Models						
GPT-5.4-High	90.2	88.6	65.5	88.1	87.5	67.4
Gemini-3.1-Pro	95.7	98.0	76.1	93.0	94.8	69.3
Qwen3-VL-Plus	87.3	88.3	66.7	87.5	89.6	70.1
Claude-4.6-Opus	89.6	84.4	53.4	89.6	89.6	59.1
Claude-4.6-Sonnet	83.9	81.2	54.5	83.6	82.3	63.6
GPT-4o-mini	41.0	38.6	29.5	41.2	40.6	29.5
Open-source Models						
Qwen3-VL-8B	61.8	63.5	37.3	55.9	64.6	32.8
InternVL3.5-8B	54.0	61.5	39.8	51.3	59.4	47.7
LLaVA-OneVision-1.5-8B	36.8	41.7	23.9	36.0	42.7	25.0
MathCoder-VL-8B	40.8	45.8	28.4	36.3	41.7	26.1
G-LLaVA-13B	16.7	30.2	25.0	17.0	32.3	23.9

Table 2: **Micro Benchmarking Results.** Per-family matched accuracy (%) aggregated over valid problem pairs. Each perturbation family is compared with Canonical on the same problem IDs; parenthesized arrows show the absolute point change and its direction. The corresponding signed effect is $\Delta_f = \widehat{\mathrm{Acc}}(F_f) - \widehat{\mathrm{Acc}}(C)$.

Model	Canonical	Orientability	Referentiality	Appearance	Redundancy
Proprietary Models					
GPT-5.4-High	84.6	80.3 (↓4.3)	83.8 (↓0.8)	82.5 (↓2.1)	86.4 (↑1.8)
Gemini-3.1-Pro	89.4	84.2 (↓5.3)	89.2 (↓0.2)	87.7 (↓1.7)	90.4 (↑1.0)
Qwen3-VL-Plus	84.9	78.5 (↓6.4)	84.7 (↓0.2)	83.1 (↓1.8)	85.0 (↑0.1)
Claude-4.6-Opus	84.5	77.8 (↓6.8)	84.0 (↓0.5)	84.9 (↑0.3)	83.6 (↓1.0)
Claude-4.6-Sonnet	80.0	73.6 (↓6.4)	77.8 (↓2.3)	80.7 (↑0.6)	80.0 (↓0.1)
GPT-4o-mini	39.1	38.1 (↓1.1)	38.3 (↓0.8)	38.3 (↓0.8)	38.8 (↓0.3)
Open-source Models					
Qwen3-VL-8B	54.5	50.4 (↓4.1)	52.3 (↓2.3)	49.9 (↓4.6)	55.3 (↑0.8)
InternVL3.5-8B	52.2	46.8 (↓5.4)	47.7 (↓4.4)	50.1 (↓2.1)	50.0 (↓2.2)
LLaVA-OneVision-1.5-8B	35.4	31.4 (↓4.0)	30.1 (↓5.3)	33.6 (↓1.8)	34.2 (↓1.2)
MathCoder-VL-8B	35.6	33.3 (↓2.3)	33.2 (↓2.4)	28.5 (↓7.1)	33.8 (↓1.8)
G-LLaVA-13B	20.9	21.8 (↑0.9)	14.3 (↓6.6)	20.2 (↓0.7)	22.1 (↑1.2)

Furthermore, to investigate the impact of domain-specific mathematical and geometric pre-training, we incorporate specialized open-source baselines: `MathCoder-VL-8B` (Wang et al., 2025a) and the geometry-focused `G-LLaVA-13B` (Gao et al., 2025).

Prompting Strategy and Metrics. All models are queried zero-shot with the same answer-format instruction and deterministic decoding. We organize the evaluation at two levels. At the *macro* level, we report Original and Canonical accuracy by source partition, treating Original-to-Canonical as a separate representation intervention. At the *micro* level, we evaluate each perturbation family on matched Canonical–variant pairs using two complementary metrics. The paired accuracy change $\Delta_f = \widehat{\mathrm{Acc}}(F_f) - \widehat{\mathrm{Acc}}(C)$ measures the average effect of transformation f on the same problem IDs, while canonical-conditioned retention $R_f^{(C)} = \widehat{\Pr}(F_f = 1 \mid C = 1)$ measures whether problems already solved in the Canonical view remain solved after the transformation.

Marginal accuracy can hide correct-to-incorrect transitions when gains on previously incorrect problems offset losses on previously solved ones; retention exposes these solved-to-failed transitions directly. Strict all-four consistency is reported separately in Appendix B.1 as a suite-level reliability measure. Answers are scored with deterministic equivalence checks and an LLM fallback, validated on 600 stratified records with 96.7% human agreement (Appendix A.4).

Table 3: **Micro Retention Results.** Canonical-conditioned family retention $R_f^{(C)}$ (%) on model-specific complete-case suites pooled over GeoQA and Geometry3K Source-Test. Complete cases have scored Canonical and all four perturbation views. Each value is the fraction of Canonical-solved suites that remain correct after the indicated transformation; all four families use the same pool within each model. Square brackets show 95% bootstrap confidence intervals computed from 10,000 resamples of the item-level predictions. Higher is better.

Model	Orient.	Refer.	Appear.	Redund.
Gemini-3.1-Pro	89.9 [83.1, 95.5]	91.0 [84.3, 96.6]	97.8 [94.4, 100.0]	94.4 [88.8, 98.9]
Claude-4.6-Sonnet	83.1 [74.0, 90.9]	93.5 [87.0, 98.7]	98.7 [96.1, 100.0]	90.9 [84.4, 96.1]
Claude-4.6-Opus	85.7 [73.8, 95.2]	88.1 [78.6, 97.6]	92.9 [83.3, 100.0]	85.7 [73.8, 95.2]
GPT-5.4-High	87.5 [77.5, 97.5]	85.0 [72.5, 95.0]	92.5 [82.5, 100.0]	92.5 [82.5, 100.0]
GPT-4o-mini	71.4 [52.4, 90.5]	61.9 [42.9, 81.0]	90.5 [76.2, 100.0]	76.2 [57.1, 95.2]

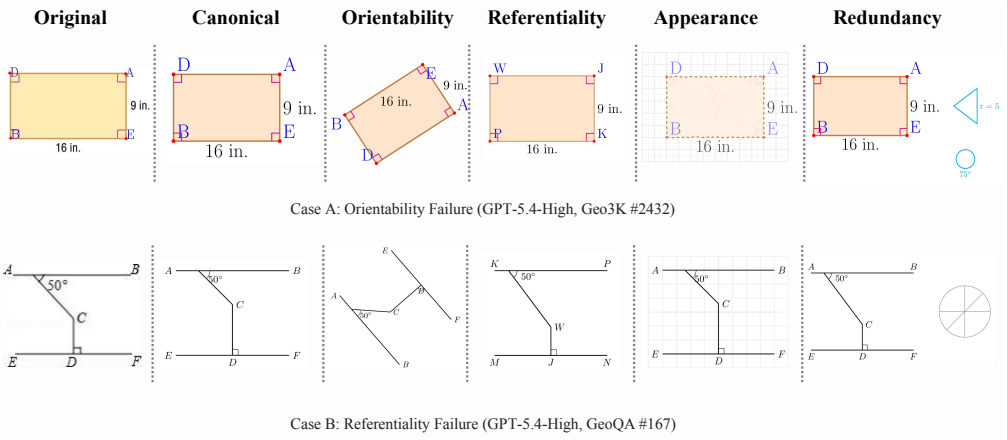

Figure 2: Representative matched failure cases under Orientability and Referentiality.

4.1 Macro Analysis: Representation-Level Effects

The primary quantitative comparison across our three source partitions is reported in Table 1.

Canonical Layout: Volatile Equivalence. A comparative audit between the source `Original` rasters and our reconstructed `Canonical` configurations reveals model-specific, non-monotonic shifts rather than a uniform trend. Because the `Canonical` view provides the shared TikZ-rendered baseline from which all perturbation variants are generated, we treat Original-to-Canonical as a separate representation intervention.

For instance, on the `Pure` split, Claude-4.6-Sonnet improves from 54.5% to 63.6%, whereas Gemini-3.1-Pro decreases from 76.1% to 69.3%. These opposite shifts show that TikZ reconstruction should not be treated as neutral preprocessing; its effect is measured separately before interpreting the downstream Canonical-to-perturbation comparisons.

4.2 Micro Analysis: Dimensional Bottlenecks

To establish a more granular understanding, we systematically deconstruct model performance across four isolated perturbation dimensions: Orientability, Referentiality, Appearance, and Redundancy. Table 2 reports matched family-level accuracy across the evaluation splits, while Table 3 reports canonical-conditioned retention on Source-Test, measuring whether Canonical-solved problems remain correct after each intervention.

Perturbed Layout: Heterogeneous Structural Sensitivity. At the micro level, the central pattern is not uniform degradation but view-dependent correctness. The same problem can be solved under one equivalent drawing and missed under another, and this instability is concentrated in specific transformations rather than generic image change. Marginal accuracy can further conceal

these failures when gains on previously incorrect problems offset losses on previously solved ones; canonical-conditioned retention exposes these solved-to-failed transitions directly (Appendix B.2). Because the `Canonical` view and all perturbation variants share the same TikZ rendering pipeline, the micro analysis separates these controlled structural interventions from the Original-to-Canonical representation shift.

On the same complete-case Source-Test pools, this cancellation is visible in the paired transition counts: 7 of the 20 model–family comparisons have non-negative net accuracy change while still containing 24 solved-to-failed transitions in total. For Claude-4.6-Sonnet under Referentiality, eight gains offset five losses, yielding a $+3.2$-point change despite 93.5% retention; for GPT-4o-mini under Redundancy, nine gains offset five losses, yielding $+8.9$ points with 76.2% retention. Thus, a stable or improving marginal score does not imply that the model preserved the problems it had already solved (Appendix B.2).

Orientability: Spatial Anchors Are Broken. Spatial orientation remains the most consistent source of degradation among the frontier models. When diagrams undergo programmatic rigid rotations, Gemini-3.1-Pro decreases from 89.4% to 84.2%, Qwen3-VL-Plus from 84.9% to 78.5%, Claude-4.6-Opus from 84.5% to 77.8%, and Claude-4.6-Sonnet from 80.0% to 73.6% relative to their matched `Canonical` baselines. More directly, Table 3 shows that rotation breaks 10.1–16.9% of problems already solved by the four strongest reported systems.

Crucially, our TikZ transformation rotates the geometric structure while keeping vertex labels upright, legible, and attached to the same entities. The observed failures persist with upright vertex labels. They expose a lack of viewpoint invariance: the model can solve the canonical arrangement yet fail to preserve the same solution after a rigid spatial transformation. An additional sweep characterizes orientation sensitivity across nine angles on both Source-Test partitions (Appendix B.5).

Referentiality: Symbolic Identities Are Fluid. Referentiality reveals a different failure pattern. At the marginal level, several strong proprietary models change little after synchronized relabeling: GPT-5.4-High, Gemini-3.1-Pro, Qwen3-VL-Plus, and Claude-4.6-Opus all change by less than one point in Table 2. The Source-Test retention analysis nevertheless reveals substantial correct-to-incorrect transitions. Among the four strongest models reported in Table 3, relabeling breaks 6.5–15.0% of Canonical-solved Source-Test problems, while GPT-4o-mini loses 38.1%.

Because the relabeling is bijective and synchronized across the diagram and text, the underlying geometry is unchanged; only the symbols naming its entities are changed. These failures therefore reveal unstable symbol-to-entity grounding. In a targeted audit, providing the original-to-current label mapping recovered only 1 of 24 failed answers (Appendix C.1). They also illustrate why marginal accuracy alone is insufficient: newly gained successes can offset previously solved problems that become incorrect after relabeling. Several open-source models show the same sensitivity directly in marginal accuracy, including LLaVA-OneVision-1.5-8B and G-LLaVA-13B, which drop by 5.3 and 6.6 points, respectively.

Appearance: Stylistic Changes Are Less Disruptive. Modifications to visual style—such as background grids, line styles, colors, or sketch-like rendering—have comparatively small effects on the strongest proprietary models. GPT-5.4-High, Gemini-3.1-Pro, and Qwen3-VL-Plus decrease by 2.1, 1.7, and 1.8 points, respectively, while canonical-conditioned retention remains above 90% for every frontier model in Table 3.

The weakness is more pronounced for some open-source models. In particular, MathCoder-VL-8B shows the largest Appearance decrease in Table 2, from 35.6% to 28.5%. Overall, the contrast with Orientability and Referentiality shows that surface style is not the main source of instability for frontier models.

Redundancy: Visual Clutter Is Mostly Neutral. Counter-intuitively, injecting non-essential visual structure produces mixed but generally small effects across models. GPT-5.4-High and Gemini-3.1-Pro improve slightly under Redundancy, whereas InternVL3.5-8B, LLaVA-OneVision-1.5-8B, and MathCoder-VL-8B show small declines.

The injected objects are detached and unlabeled, with no proof-relevant relation to the core construction. Their limited effect therefore indicates that irrelevant visual content itself is not the main source

of structural instability. Instead, Redundancy primarily tests whether models can selectively ignore information that is unnecessary for the solution.

The Failure Profile Is Selective, Not Generic. Taken together, the four families reveal selective sensitivity: rotation produces the most consistent net loss, while relabeling also breaks previously solved problems. GeoPerturb therefore does not merely show that image changes hurt performance; it identifies which invariances fail. Appearance changes and irrelevant clutter generally have smaller net effects for frontier models.

A per-set decomposition further shows that most Canonical-solved problems retain correctness across multiple perturbation families. Across the five evaluation sets and five reported models, $k = 4$ is a modal outcome in 24 of 25 model–set cells; for the four strongest systems, failures on three or four families are uncommon and the mean number of retained families ranges from 3.25 to 3.87 (Appendix B.3).

4.3 Case Analysis

To make the aggregate diagnostics concrete, we trace two representative matched failures in Figure 2:

- **Orientability Inconsistency:** In Case A (#2432), GPT-5.4-High correctly solves the canonical rectangle-area problem (16×9) but fails after a rigid rotation that preserves the side lengths, right angles, labels, and gold answer. The model changes the reported width from 9 to 6 while retaining the area calculation (Appendix F). It directly demonstrates viewpoint sensitivity on this item.
- **Referentiality Inconsistency:** In Case B (#167), GPT-5.4-High correctly resolves the problem under the standard labels (A, B, C) but fails after a synchronized bijective remapping (K, P, W). Every label change is synchronized across the diagram and text, so the underlying geometry and intended solution remain unchanged. The correct-to-incorrect transition shows that the model does not preserve the updated symbol-to-entity mapping through the solution.

Together, these cases make the retention result concrete: a problem solved in the `Canonical` view can become incorrect after an answer-preserving representation change, particularly under the rotation and relabeling families that expose the clearest aggregate instability.

4.4 Perturbation-guided Finetuning

The fine-grained diagnostic capacity of GeoPerturb naturally motivates a critical downstream question: if layout perturbations expose structural vulnerabilities, can targeted supervision actively mitigate these failures? To investigate this intervention potential, we conduct a controlled finetuning study on the open-source `Qwen3-VL-8B-Instruct` baseline. We separate the training view (`Original`, `Canonical`, or `Perturbations`) from the supervision target (derivation plus answer, or answer only). Derivation-supervised conditions are trained to produce an explicit derivation followed by the answer, whereas AnswerOnly conditions are trained to output only the answer. Crucially, all conditions—Base, derivation-supervised, and AnswerOnly—are evaluated on the same Geometry3K Source-Test items and view sets with the same CoT-formatted evaluation prompt. Correctness is judged only by final-answer equivalence, so a model is not penalized for omitting a derivation. "Avg. perturbation" is the arithmetic mean over the four family accuracies.

Holding the perturbation training view fixed, `Perturb-CoT` exceeds `Perturb-AnswerOnly` by 38.8, 49.0, and 42.6 points on Original, Canonical, and average perturbation accuracy, respectively, and exceeds the Base model by 8.2, 8.2, and 11.9 points. The same supervision-target contrast persists across training views: derivation-supervised conditions outperform answer-only conditions by 34.7–49.0 points on Canonical and 37.7–42.6 points on average perturbation accuracy for `Perturbations`, `Canonical TikZ`, and `Original raster` training alike. Under the same evaluation prompt, `Perturb-CoT` also achieves higher average perturbation accuracy (72.0%) than `Original-CoT` (68.5%) and `Canonical-CoT` (66.4%).

The controlled comparison yields a clear conclusion: more views alone are not enough. Answer-only finetuning remains weak even when trained on the same perturbed examples, whereas explicit derivation supervision yields substantially stronger cross-view performance. For `Qwen3-VL-8B-Instruct`,

Table 4: Controlled finetuning results on Geometry3K Source-Test (%). All rows use the same CoT-formatted evaluation prompt and answer-equivalence scorer. The key matched comparison, `Perturb-CoT` versus `Perturb-AnswerOnly`, uses the same perturbation training data and differs only in the supervision target.

Condition	Training view	Supervision	Original	Canonical	Avg. perturbation
Base	–	–	61.2	71.4	60.1
Perturb-CoT	Perturbations	Derivation + answer	69.4	79.6	72.0
Original-CoT	Original raster	Derivation + answer	77.6	73.5	68.5
Canonical-CoT	Canonical TikZ	Derivation + answer	71.4	71.4	66.4
Perturb-AnswerOnly	Perturbations	Answer only	30.6	30.6	29.4
Canonical-AnswerOnly	Canonical TikZ	Answer only	30.6	36.7	28.7
Original-AnswerOnly	Original raster	Answer only	32.7	34.7	27.3

the useful intervention is therefore not perturbation exposure by itself, but training that teaches an explicit solution process across those representations.

A separate rendering-source control further shows that matched AnswerOnly runs trained on Canonical TikZ renders versus Original rasters differ by at most 2.5 points across five evaluation sets, with all paired confidence intervals overlapping zero (Appendix G.1). This additional control shows that changing the training-image source alone has a much smaller effect than changing the supervision target in Table 4.

5 Conclusion

In this work, we have challenged the prevailing evaluation paradigms of LMMs in geometric reasoning by introducing GeoPerturb, a programmatic framework that tests whether geometric solutions remain stable across answer-preserving redraws. Through standardized TikZ reconstructions and four isolated perturbation families, we conduct a matched diagnostic evaluation across proprietary and open-source models.

Our empirical findings show that correctness can be a property of the problem–view pair rather than a stable property of the problem. This instability is selective rather than generic: rotation and relabeling expose the clearest weaknesses, while appearance changes and irrelevant visual structure are less disruptive for frontier models. Marginal accuracy can further conceal these failures when gains on previously incorrect problems offset losses on problems already solved; canonical-conditioned retention exposes these solved-to-failed transitions directly. A controlled fine-tuning study further shows that derivation-supervised finetuning on perturbed views improves cross-view performance, whereas answer-only finetuning on the same views does not. By shifting benchmark evaluation from static single-view accuracy to matched within-problem invariance, GeoPerturb offers a diagnostic lens for asking whether a model preserves the same solution across equivalent representations. Solving one drawing is not the same as solving the geometry.

Limitations

GeoPerturb measures representation sensitivity through end-to-end model behavior. Localizing a failure to the vision encoder, cross-modal interface, or reasoning stack requires additional internal probes. Semantic validity is established through independent human review and correction rather than formal certification (Appendix A). The `Pure` split contains 88 independently collected problems; its public source material may overlap undisclosed proprietary pretraining data. The fine-tuning results are established for `Qwen3-VL-8B-Instruct`.

The current perturbation engine covers static 2D Euclidean geometry. Extending GeoPerturb to 3D, dynamic, or non-Euclidean settings requires new representations and validity-preserving transformation rules.

AI USE DISCLOSURE

Generative AI tools were used in a limited assistive capacity for language polishing, routine coding and debugging, producing candidate TikZ renderings and perturbation implementations during data preparation, and synthesizing chain-of-thought (CoT) training data for fine-tuning. The research questions, conceptual framework, perturbation definitions, dataset curation, validation procedures, experiments, analyses, and conclusions were designed and supervised by the authors. All AI-assisted outputs were reviewed, tested, corrected, and accepted only after author verification. AI tools were not used to generate the research ideas, draft manuscript sections, prove mathematical claims, or determine the final scientific conclusions. The authors take full responsibility for the final content of this paper.

Separately from this assistive use, language models are components of the released method rather than tools used to produce the manuscript: a Gemini-Flash model generates the Canonical TikZ reconstructions and applies the perturbation edits, a Gemini-Flash model generates the reasoning traces used in the fine-tuning study (Appendix G), and an LLM fallback judge supplies the evaluation verdict whenever deterministic matching is inconclusive (Appendix A.4).

REFERENCES

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Anthropic. Claude opus 4.6 system card, 2026a. URL `https://www.anthropic.com/claude-opus-4-6-system-card`.

Anthropic. Claude sonnet 4.6 system card, 2026b. URL `https://www.anthropic.com/claude-sonnet-4-6-system-card`.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Jonas Belouadi, Anne Lauscher, and Steffen Eger. Automatikz: Text-guided synthesis of scientific vector graphics with tikz. In *International Conference on Learning Representations*, volume 2024, pp. 55917–55943, 2024a.

Jonas Belouadi, Simone P Ponzetto, and Steffen Eger. Detikzify: Synthesizing graphics programs for scientific figures and sketches with tikz. *Advances in Neural Information Processing Systems*, 37: 85074–85108, 2024b.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 513–523, 2021.

Luoxin Chen, Jinming Gu, Liankai Huang, Wenhao Huang, Zhicheng Jiang, Allan Jie, Xiaoran Jin, Xing Jin, Chenggang Li, Kaijing Ma, et al. Seed-prover: Deep and broad reasoning for automated theorem proving. *arXiv preprint arXiv:2507.23726*, 2025.

Seunghyuk Cho, Zhenyue Qin, Yang Liu, Youngbin Choi, Seungbeom Lee, and Dongwoo Kim. Plane geometry problem solving with multi-modal reasoning: A survey. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 110–131, 2026.

Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24625–24634, 2024.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76 (5):378–382, 1971. doi: 10.1037/h0031619.

Deqing Fu, Ruohao Guo, Ghazal Khalighinejad, Ollie Liu, Bhuwan Dhingra, Dani Yogatama, Robin Jia, and Willie Neiswanger. Isobench: Benchmarking multimodal foundation models on isomorphic representations. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=KZd1EErRJ1.

Yumeng Fu, Jiayin Zhu, Lingling Zhang, Bo Zhao, Shaoxuan Ma, Yushun Zhang, Yanrui Wu, and Wenjun Wu. Geolaux: A benchmark for evaluating mllms' geometry performance on long-step problems requiring auxiliary lines. *arXiv preprint arXiv:2508.06226*, 2025.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. In *International Conference on Learning Representations*, pp. 3490–3511, 2025.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Google DeepMind. Gemini 3.1 pro model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Christian Greisinger and Steffen Eger. Tikzilla: Scaling text-to-tikz with high-quality data and reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=rJv2byEWA3.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Yangyi Li, Chenxu Zhao, and Mengdi Huai. Quantifying and understanding uncertainty in large reasoning models. *arXiv preprint arXiv:2604.13395*, 2026.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6774–6786, 2021.

OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence, 2024. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

OpenAI. Introducing gpt-5.4, 2026. URL https://openai.com/index/introducing-gpt-5-4/.

Bowen Ping, Minnan Luo, Zhuohang Dang, Chenxi Wang, and Chengyou Jia. AutoGPS: Automated geometry problem solving via multimodal formalization and deductive reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=PVtZnUh04m.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 21527–21536, 2024.

Jielin Qiu, Yi Zhu, Xingjian Shi, Zhiqiang Tang, Ding Zhao, Bo Li, and Mu Li. Benchmarking robustness under distribution shift of multimodal image-text models. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022. URL https://openreview.net/forum?id=PN4CUHbnlc.

Itsumi Saito, Haruto Yoshida, and Keisuke Sakaguchi. Sketch2diagram: Generating vector diagrams from hand-drawn sketches. In *International Conference on Learning Representations*, volume 2025, pp. 14332–14354, 2025.

Nannan Shi, Chuanyu Qin, Shipeng Song, and Man Luo. Geothought: A dataset for enhancing mathematical geometry reasoning in vision-language models. *arXiv preprint arXiv:2510.21881*, 2025.

Jiayin Sun, Caixia Sun, Boyu Yang, Hailin Li, Xiao Chen, Yi Zhang, Errui Ding, Liang Li, Chao Deng, and Junlan Feng. Geotikzbridge: Advancing multimodal code generation for geometric perception and reasoning. *arXiv preprint arXiv:2603.22687*, 2026.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840*, 2021.

Ke Wang, Junting Pan, Linda Wei, Aojun Zhou, Weikang Shi, Zimu Lu, Han Xiao, Yunqiao Yang, Houxing Ren, Mingjie Zhan, and Hongsheng Li. Mathcoder-vl: Bridging vision and code for enhanced multimodal mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 2505–2534, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Xiaojing Wei, Ting Zhang, Wei He, Jingdong Wang, and Hua Huang. Geoloom: High-quality geometric diagram generation from textual input. *arXiv preprint arXiv:2512.08180*, 2025.

Chen Wu and Chao Sun. Structured reasoning for geometry problems: A chain-of-knowledge approach with structurally retrieved examples. In *2025 International Conference on Intelligent Education and Intelligent Research (IEIR)*, pp. 1–9. IEEE, 2025.

Shihao Xu, Yiyang Luo, and Wei Shi. Geo-llava: A large multi-modal model for solving geometry math problems with meta in-context learning. In *Proceedings of the 2nd Workshop on Large Generative Models Meet Multimodal Applications*, pp. 11–15, 2024.

Liang Yan, Xu Jiang, Jian Ma, Yuhang Liu, Tian Bian, Qichao Wang, Abhishek Basu, Yu Rong, Tingyang Xu, Pengcheng Wu, et al. A comprehensive survey of multimodal llms for scientific discovery. In *1st Workshop on VLM4RWD@ NeurIPS 2025*, 2025.

Jiacheng Yang, Anqi Chen, Yunkai Dang, Qi Fan, Cong Wang, Wenbin Li, Feng Miao, and Yang Gao. Annotation-free visual reasoning for high-resolution large multimodal models via reinforcement learning. *arXiv preprint arXiv:2602.23615*, 2026a.

Tianyu Yang, Sihong Wu, Yilun Zhao, Zhenwen Liang, Lisen Dai, Chen Zhao, Minhao Cheng, Arman Cohan, and Xiangliang Zhang. A survey of multimodal mathematical reasoning: From perception, alignment to reasoning. *arXiv preprint arXiv:2603.08291*, 2026b.

Yufei Zhan, Hongyin Zhao, Yousong Zhu, Shurong Zheng, Fan Yang, Ming Tang, and Jinqiao Wang. Understand, think, and answer: Advancing visual reasoning with large multimodal models. *arXiv preprint arXiv:2505.20753*, 2025.

Chengrui Zhang, Maizhen Ning, Tianyi Liu, Zihao Zhou, Jie Sun, Qiufeng Wang, and Kaizhu Huang. Geosdf: Plane geometry diagram synthesis via signed distance field. *arXiv preprint arXiv:2506.13492*, 2025.

Yurui Zhao, Xiang Wang, Jiahong Liu, Irwin King, and Zhitao Huang. Towards geometry problem solving in the large model era: A survey. *arXiv preprint arXiv:2506.02690*, 2025.

Hao Zhong, Muzhi Zhu, Zongze Du, Zheng Huang, Canyu Zhao, Mingyu Liu, Wen Wang, Hao Chen, and Chunhua Shen. Omni-R1: Reinforcement learning for omnimodal reasoning via two-system collaboration. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 45266–45297, 2025.

Kaijie Zhu, Qinlin Zhao, Hao Chen, Jindong Wang, and Xing Xie. Promptbench: A unified library for evaluation of large language models. *Journal of Machine Learning Research*, 25(254):1–22, 2024.

Zhihao Zhu, Jiafeng Liang, Shixin Jiang, Jinlan Fu, Ming Liu, Guanglu Sun, See-Kiong Ng, and Bing Qin. Analyzing reasoning consistency in large multimodal models under cross-modal conflicts. *arXiv preprint arXiv:2601.04073*, 2026.

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. In *International Conference on Learning Representations*, 2025.

A Dataset Construction and Validation

This appendix documents the two checks underlying matched evaluation: whether each constructed view preserves the problem and its intended answer, and whether model responses are scored correctly.

A.1 Coverage Funnel

Table 5 reports 531 matched problem suites from 553 candidate problems (96.0% coverage). We report reconstruction coverage explicitly because TikZ reconstruction is treated as a measured representation intervention rather than hidden preprocessing.

Table 5: Coverage funnel. "Matched" denotes problems retained for paired original–canonical and canonical–perturbation analyses.

Subset	Candidate	Reconstructed	Coverage	Matched
GeoQA Source-Test	49	47	95.9%	47
GeoQA Source-Train	164	155	94.5%	155
Geometry3K Source-Test	51	49	96.1%	49
Geometry3K Source-Train	199	192	96.5%	192
Pure	90	88	97.8%	88
Total	553	531	96.0%	531

A.2 Human Audit and Manual Correction

We used the same review procedure for every rendered view. Three graduate students in mathematics education independently labeled each view as VALID, INVALID, or UNCERTAIN. The review checked whether the rendered diagram preserved the original problem and its answer. We corrected every non-VALID view and re-audited it until it passed. The released dataset contains the post-correction views. Seventy-three reconstructions that failed the initial automatic pass were manually corrected and re-audited rather than discarded (Table 7).

A.3 Inter-Annotator Agreement

We treated semantic validity as a human judgment. To measure the consistency of this judgment, three graduate students in mathematics education independently labeled every view in the release. The release contains 531 Canonical views and 2,124 perturbation variants. Each annotator saw the source and target renders, the question, the gold answer, and the intended transformation. They did not see model outputs, scorer decisions, or the other annotators' labels. Each view received one label: VALID, INVALID, or UNCERTAIN.

We computed agreement on the pre-adjudication labels using Fleiss' multi-rater κ (Fleiss, 1971). Across all 2,655 views, Fleiss' $\kappa = 0.87$ and raw agreement was 98.6%. Here, raw agreement is the proportion of views for which all three annotators assigned the same label. Table 8 reports the agreement for each view type. The κ values range from 0.84 to 0.91 across the five view types. We report both measures because most views are valid: this makes raw agreement high and also increases chance agreement, which lowers κ.

Across the 7,965 individual labels, 297 were non-VALID: 56 INVALID and 241 UNCERTAIN. Every non-VALID view was corrected and re-audited until it passed. The retained release contains the post-correction views, so these 297 labels describe pre-adjudication annotator behavior rather than defects in the released set.

A.4 Scoring Validation

All final results use a rule-first hybrid scorer. The deterministic stage applies exact and normalized string matching, numerical matching, and SymPy equivalence; when these rules are inconclusive, an LLM fallback judges mathematical equivalence. The fallback judge is the evaluated `GPT-4o-mini`. The independent cross-check judge is text-only `DeepSeek-V3.2`; it is used only for validation and is not part of the evaluated model set. Both judges use the same prompt verbatim at temperature 0

Table 6: Semantic audit criteria used for canonical reconstructions and perturbation variants.

Family	Accepted	Rejected or corrected
General legibility	Minor label–line proximity or overlap is allowed when the symbol and its ownership remain unambiguous under human inspection and automated checks.	Clipping, occlusion, unreadable symbols, or ambiguous label ownership.
Orientability	The geometric structure is rigidly transformed; labels remain upright, legible, and attached to the correct entities.	Rotated text, detached labels, clipping, or changed primitives/incidences.
Referentiality	A one-to-one relabeling is synchronized across the diagram, question, options, and answer representation when needed.	Collisions, incomplete mappings, or inconsistent text–diagram labels.
Appearance	Grid, color, dashed-line, or sketch styling may alter pixels and create background crossings, while primitives, labeled relations, constraints, and the answer remain unchanged.	A style change obscures a required primitive or introduces a proof-relevant mark or constraint.
Redundancy	Detached, unlabeled auxiliary objects may share the canvas and be visually nearby while remaining separate from the answer-critical construction.	Contact, intersection, tangency, collinearity, labeling, or another apparent constraint involving the core geometry.

Table 7: Manual reconstruction corrections by subset. These are diagrams that failed the initial automatic reconstruction and were corrected before inclusion.

Subset	Corrected diagrams
GeoQA Source-Test	12
GeoQA Source-Train	9
Geometry3K Source-Test	7
Geometry3K Source-Train	25
Pure	20
Total	73

with a 16-token binary output budget. We validated the scoring pipeline on a stratified sample of 600 records. A human annotator agreed with the automatic verdict on 580 records (96.7%, Cohen's $\kappa = 0.93$). The cross-check judge agreed on 578 records (96.3%, $\kappa = 0.92$).

The fallback judge is itself one of the evaluated models. Agreement was 58/60 for `GPT-4o-mini` outputs and 520/540 for the other outputs ($p \approx 0.88$). Agreement rates were similar for the two groups.

Agreement across view types ranged from 93.3% to 99.2%. A similar ordering appeared across the five view types in the semantic-validity audit (Spearman $\rho = 0.90$). We use this correlation as an ordering check rather than as a powered statistical test.

The validation sample also contained 22 paired Canonical–perturbation transitions. Human review confirmed the direction of every transition. Both judges saw only the problem, the gold answer, and the model response; they did not see the view type or the model identity. Across these checks, we found no substantial evidence of view-dependent scoring bias. The paired checks also make it unlikely that the fallback scorer alone explains the direction of Δ_f.

The exact judging prompt appears in Appendix H.4. Normalization rules and numerical-tolerance settings are released with the scorer.

B Supplementary Robustness Summary

The main paper uses per-family paired effects and canonical-conditioned retention as the primary perturbation metrics. This section retains strict all-four consistency as a complementary suite-level reliability diagnostic and preserves the compact delta table to make the direction of each perturbation family explicit. The deltas in Table 12 use the same matched aggregate as Table 2. Since the main table rounds accuracies to one decimal, some deltas are computed before rounding.

Table 8: Inter-annotator agreement on the three-way semantic validity judgment (VALID / INVALID / UNCERTAIN), computed on pre-adjudication labels. All 2,655 views were annotated independently by three graduate students in mathematics education.

View type	Items	Fleiss' κ
Canonical reconstruction	531	0.87
Orientability (spatial rotation)	531	0.90
Referentiality (symbol permutation)	531	0.91
Appearance (visual stylization)	531	0.84
Redundancy (irrelevant injection)	531	0.85
All views	2,655	0.87

B.1 Suite-Level Consistency

For a complete suite with four perturbation families, we define canonical-conditioned all-four retention and the unconditional strict joint score as

$$R_{4|C} = \frac{\sum_i C_i \prod_{f \in \mathcal{F}} F_{i,f}}{\sum_i C_i},$$

$$J = \frac{1}{n} \sum_i \prod_{f \in \mathcal{F}} F_{i,f}.$$

$R_{4|C}$ asks whether a problem solved in the Canonical view remains correct across all four answer-preserving variants, while J is the unconditional pass rate of the complete four-variant suite. Both are stricter reliability criteria than single-view accuracy. Neither is interpreted as a perturbation-induced accuracy drop, and neither is subtracted from Canonical accuracy.

Table 9: Canonical-conditioned all-four retention $R_{4|C}$ (%) pooled over GeoQA and Geometry3K Source-Test. Square brackets show 95% bootstrap confidence intervals computed from 10,000 resamples of the item-level predictions.

Model	$R_{4\|C}$
Gemini-3.1-Pro	76.4 [67.4, 84.3]
Claude-4.6-Sonnet	72.7 [62.3, 81.8]
Claude-4.6-Opus	69.0 [54.8, 81.0]
GPT-5.4-High	65.0 [50.0, 80.0]
GPT-4o-mini	33.3 [14.3, 52.4]

Table 10: Strict joint score J (%) for complete four-variant suites. This table reports descriptive suite-level reliability rather than a Canonical-minus-Perturbed effect. N is the complete-case denominator for each split.

	Source-Train		Source-Test		Pure	
Model	J	N	J	N	J	N
Proprietary Models						
GPT-5.4-High	71.5	344	60.0	45	51.2	80
Gemini-3.1-Pro	80.9	345	75.2	93	50.0	88
Qwen3-VL-Plus	67.3	345	70.3	64	42.4	85
Claude-4.6-Opus	71.6	342	66.7	45	30.7	88
Claude-4.6-Sonnet	62.6	345	61.3	93	39.8	88
GPT-4o-mini	20.6	345	17.8	45	14.0	86
Open-source Models						
Qwen3-VL-8B	27.6	344	31.1	45	15.2	66
InternVL3.5-8B	23.5	344	26.7	45	25.0	88
LLaVA-OneVision-1.5-8B	15.1	344	24.4	45	10.2	88
MathCoder-VL-8B	16.6	344	22.2	45	12.5	88
G-LLaVA-13B	4.7	344	13.3	45	9.1	88

To distinguish genuine cross-family dependence from the mechanical reduction caused by requiring four simultaneous successes, we construct a marginal-preserving permutation reference within Canonical-correct Pure suites. Each family's correctness vector is independently permuted, preserving its conditional retention rate while removing cross-family dependence. As shown in Table 11, the observed all-four retention exceeds this reference by 8.9–18.3 percentage points across all five reported models, indicating that the strict criterion contains dependence information beyond the four marginal retention rates.

Table 11: Pure split all-four retention versus the marginal-preserving permutation reference (%).

Model	Observed	Reference	Difference
Gemini-3.1-Pro	70.5	60.9	$+9.6$
Claude-4.6-Sonnet	62.5	44.2	$+18.3$
Claude-4.6-Opus	51.9	43.0	$+8.9$
GPT-5.4-High	74.1	59.3	$+14.8$
GPT-4o-mini	48.0	36.4	$+11.6$

Extending the same marginal-preserving reference across all five evaluation sets, observed all-four retention exceeds the reference in 23 of 25 model–set cells. After Holm adjustment, nine excesses remain significant and no cell shows a significant deficit, indicating that failures tend to cluster on a subset of problems rather than occurring independently across perturbation families.

B.2 Solved-to-Failed Decomposition

Figure 3 separates lost solutions from newly gained ones on the model-specific complete-case Source-Test pools used in Table 3. A complete case has scored Canonical and all four perturbation views. For each model, let N be the pool size and n_C the number of Canonical-correct problems. For family f, let L_f count Canonical-correct problems that become incorrect and G_f count the reverse transitions. Then

$$\Delta_f = \frac{G_f - L_f}{N}, \qquad R_f^{(C)} = 1 - \frac{L_f}{n_C}.$$

Seven of the 20 model–family comparisons have non-negative net accuracy change, yet together contain 24 solved-to-failed transitions, counted at the model–problem–family level. For Claude-4.6-Sonnet under Referentiality, eight gains offset five losses: accuracy increases by 3.2 percentage points while retention is 93.5%. For GPT-4o-mini under Redundancy, nine gains offset five losses, yielding a +8.9-point change with 76.2% retention. Unchanged or improved marginal accuracy can therefore coexist with the loss of previously solved problems.

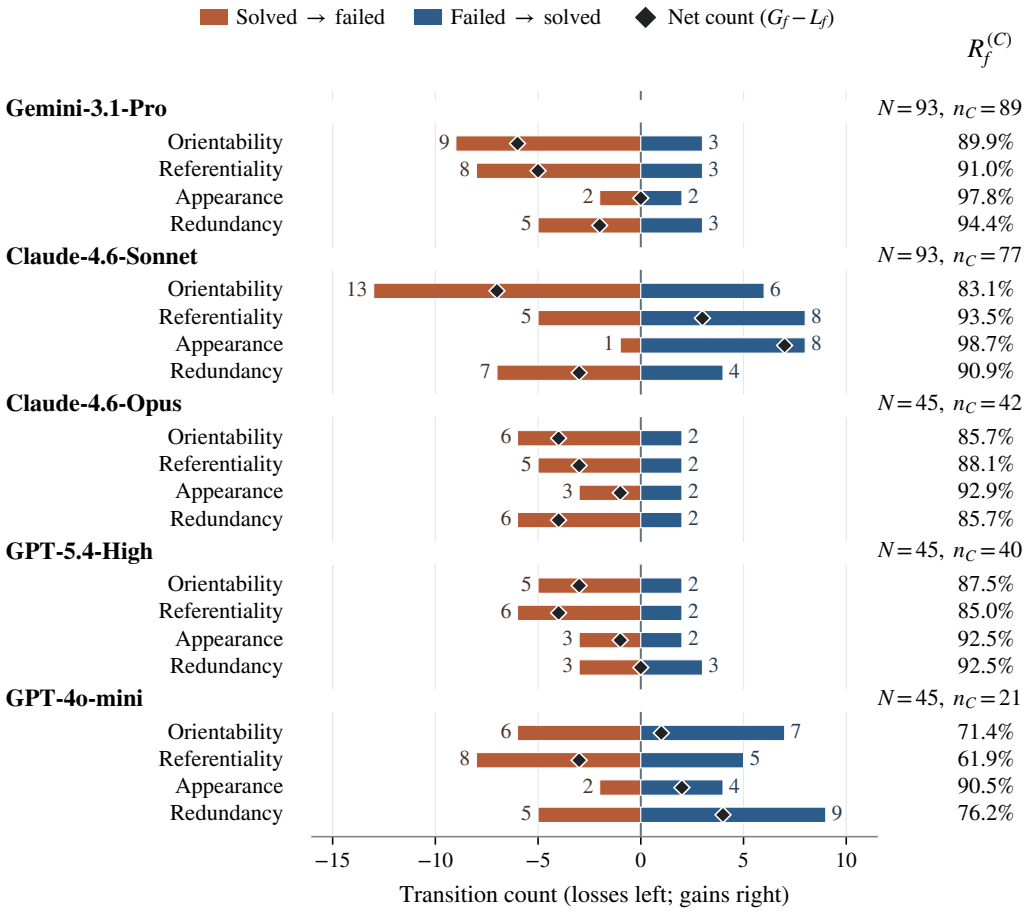

Figure 3: **Paired correctness transitions on Source-Test.** Leftward bars count solved-to-failed transitions (L_f), and rightward bars count failed-to-solved transitions (G_f). Black diamonds show the net count $G_f - L_f$; dividing by N and multiplying by 100 gives the accuracy change in percentage points. Each model's header reports N and n_C. The right-hand column reproduces the conditional-retention point estimates in Table 3.

B.3 Depth of Conditional Failure

For each Canonical-correct complete suite, let $k \in \{0, \ldots, 4\}$ denote the number of perturbation families that remain correct. Figure 4 shows the distributions on the two Source-Test partitions and Pure. Across all five evaluation sets and five reported models, $k = 4$ is a modal outcome in 24 of 25 model–set cells, including a tie with $k = 3$ for GPT-4o-mini on GeoQA Source-Test. The only cell where $k = 4$ is not a mode is GPT-4o-mini on Geometry3K Source-Test, where $k = 3$ is most common. Deep failures with $k \leq 1$ are rare for the four strongest systems, ranging from 0.0% to 8.9% across sets, and their mean retained-family count ranges from 3.25 to 3.87; for GPT-4o-mini, the corresponding mean ranges from 2.85 to 3.16.

Thus, among Canonical-solved problems, correctness is usually retained across multiple views; failures on three or four families are uncommon for the four strongest reported systems.

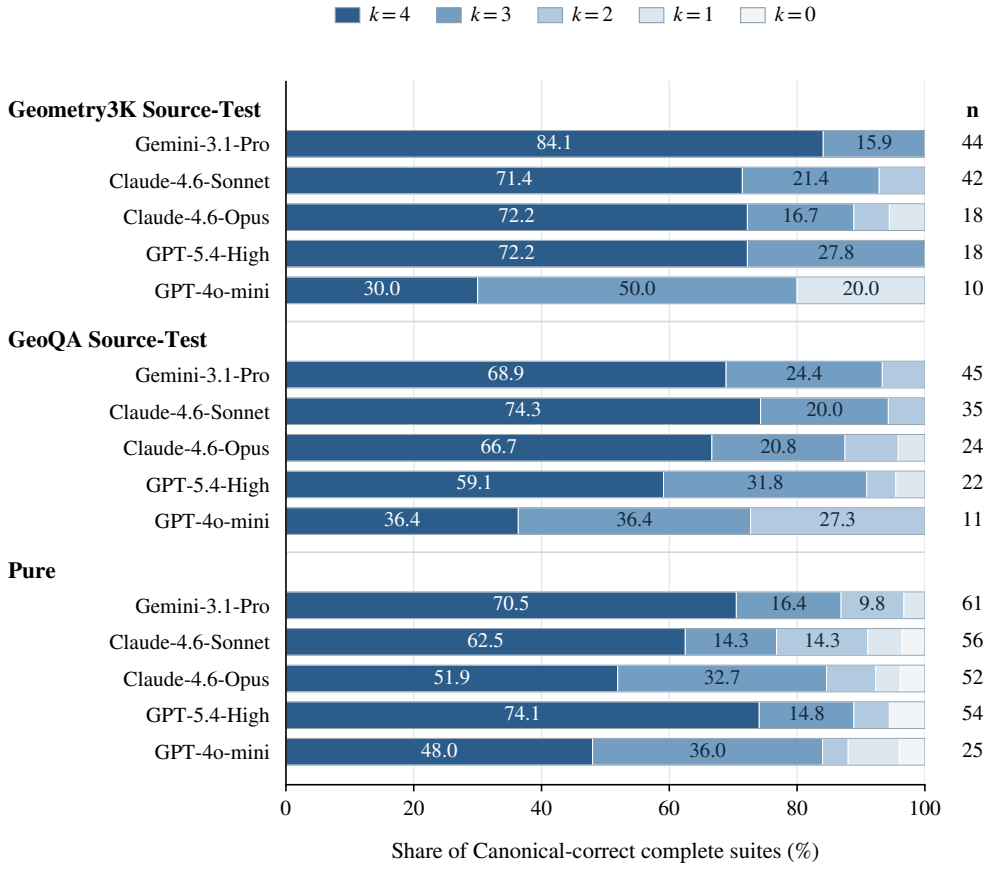

Figure 4: **Conditional failure depth.** Each bar shows the distribution of k, the number of perturbation families answered correctly, among Canonical-correct complete suites. The dark-blue $k = 4$ segment equals all-four retention $R_{4|C}$; the rightmost column gives the denominator n. Labels inside sufficiently large segments show percentages. The figure displays 15 Source-Test/Pure model–set cells; the 25-cell summary in the text additionally includes both Source-Train partitions.

B.4 Dimension-Level Accuracy Deltas

For each perturbation family f, we report

$$\Delta_f = \widehat{\mathrm{Acc}}(F_f) - \widehat{\mathrm{Acc}}(C).$$

Negative values indicate accuracy loss relative to the canonical reconstruction, while positive values indicate accuracy gain.

Table 12: Per-dimension accuracy delta relative to Canonical (percentage points). Negative means the perturbation hurts; positive means it helps.

Model	Orientability	Referentiality	Appearance	Redundancy
Proprietary Models				
GPT-5.4-High	-4.3	-0.8	-2.1	$+1.8$
Gemini-3.1-Pro	-5.3	-0.2	-1.7	$+1.0$
Qwen3-VL-Plus	-6.4	-0.2	-1.8	$+0.1$
Claude-4.6-Opus	-6.8	-0.5	$+0.3$	-1.0
Claude-4.6-Sonnet	-6.4	-2.3	$+0.6$	-0.1
GPT-4o-mini	-1.1	-0.8	-0.8	-0.3
Open-source Models				
Qwen3-VL-8B	-4.1	-2.3	-4.6	$+0.8$
InternVL3.5-8B	-5.4	-4.4	-2.1	-2.2
LLaVA-OneVision-1.5-8B	-4.0	-5.3	-1.8	-1.2
MathCoder-VL-8B	-2.3	-2.4	-7.1	-1.8
G-LLaVA-13B	$+0.9$	-6.6	-0.7	$+1.2$

B.5 Rotation-Angle Sweep

Figure 5 reports an additional rotation sweep for Gemini-3.1-Pro, GPT-5.4-High, and GPT-4o-mini at $0°$, $15°$, $30°$, $45°$, $60°$, $90°$, $120°$, $150°$, and $180°$. Each angle has 49 scored problems on Geometry3K Source-Test and 47 on GeoQA Source-Test. Accuracy varies non-monotonically with angle. For example, GPT-4o-mini on Geometry3K falls from 36.73% at $0°$ to 18.37% at $45°$, then rises to 34.69% at $60°$. On GeoQA, its $0°$ and $180°$ accuracies are both 31.91%, while intermediate-angle accuracies range from 23.40% to 36.17%. These trajectories show why an endpoint comparison alone can miss orientation sensitivity.

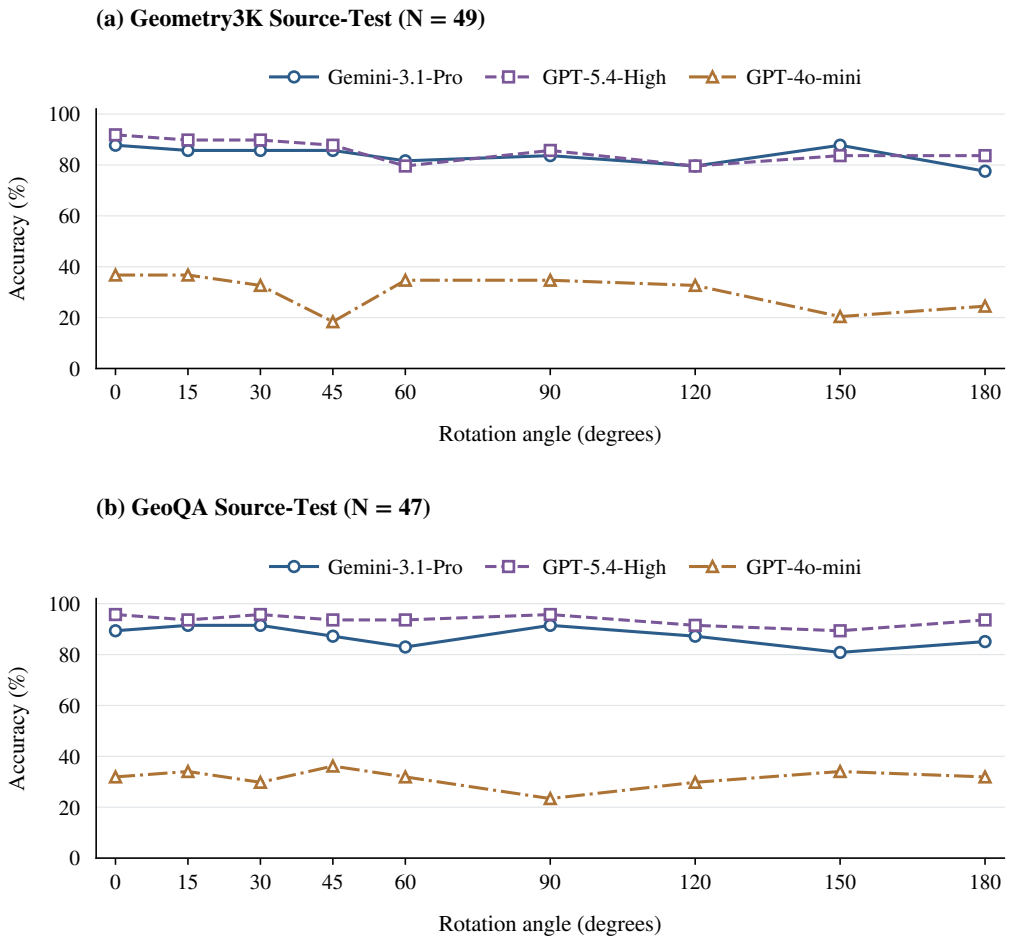

Figure 5: **Rotation-angle sensitivity.** Accuracy at nine evaluated angles for three models. Panels show Geometry3K Source-Test and GeoQA Source-Test; N is the number of scored problems at each angle. Markers identify the sampled angles, and connecting lines show the trajectories between them; the annotated Δ is the signed $0° \to 180°$ change for that model. The $0°$ value is the reference within each sweep.

C Referential Grounding Diagnostics

Referentiality is implemented through synchronized label remapping. In this appendix we refer to this implementation as SymbolMapping. It preserves the underlying geometry while changing the label-to-entity bindings, and therefore provides a direct probe of symbolic grounding.

C.1 Mechanism Audit: Perception vs. Binding

To separate label perception from symbolic binding, we conduct a targeted audit on Referentiality failures. A failure case is a model–problem pair where the model answers the canonical reconstruction correctly but answers the Referentiality version incorrectly. We audit 24 such model–problem pairs.

Label-reading probe. For each failure case, the model is shown the Referentiality diagram and asked to list all visible point labels. The reported labels are compared against labels extracted from the perturbed TikZ code. Fourteen responses satisfy the probe's validity criteria; across these responses, mean label recall is 98.0% and mean precision is 88.6%. The relabeled symbols are therefore usually recoverable in the 14 valid probe responses.

Oracle mapping recovery. For the same 24 failure cases, we provide an explicit original-to-current label mapping table before solving. Only 1 of 24 failures recovers the correct answer (4.2%). Thus, even when the remapping is made explicit, most failures persist.

Interpretation. Across the 14 valid label-reading responses, mean label recall is 98.0%; providing the label mapping recovers only 1 of the 24 audited failures. The results are consistent with unstable downstream symbol use or multi-step binding. This targeted audit does not identify a unique internal mechanism or estimate a population-wide failure rate.

Complementary Orientability control. As a complementary check, we compare rotations in which labels rotate with the diagram against the upright-label variant used by GeoPerturb on Geometry3K Source-Test ($N = 49$). For GPT-5.4-High, the mean rotation drop decreases from 6.9 to 3.6 points when labels are kept upright; for GPT-4o-mini, the corresponding drops are 7.4 and 8.4 points. Keeping vertex labels upright does not eliminate the observed rotation sensitivity.

D Split Interpretation and Template-Reliance Screen

Source-Train–Source-Test comparisons cannot prove data contamination because the partitions can differ in difficulty, rendering style, and diagram distribution. We therefore treat perturbation asymmetry only as a behavioral screen: a larger Source-Train advantage under answer-preserving interventions is consistent with presentation familiarity, but it is not forensic evidence of pretraining membership.

The 88 `Pure` problems were independently collected from public advanced-geometry examination material, OCR-processed, cleaned, normalized, and screened against GeoQA, Geometry3K, MathVista, and other mainstream multimodal benchmark pools. This excludes direct sampling from the reused benchmark pools by construction, but it cannot exclude the possibility that an undisclosed proprietary pretraining corpus contained an original public examination page. We therefore treat `Pure` as an independently constructed out-of-distribution evaluation split rather than as contamination-free. Accordingly, split-level comparisons are interpreted as robustness diagnostics rather than as direct claims about training-data membership.

E Statistical Calibration

All paired comparisons are computed over matched problem suites. For reconstruction effects, comparisons are paired by problem between the Original and Canonical views and evaluated using exact paired McNemar tests. For perturbation-family effects, comparisons are paired between the Canonical view and the corresponding family-specific perturbation. The paired effects Δ_f and canonical-conditioned retention estimates use 10,000 problem-level paired bootstrap resamples to

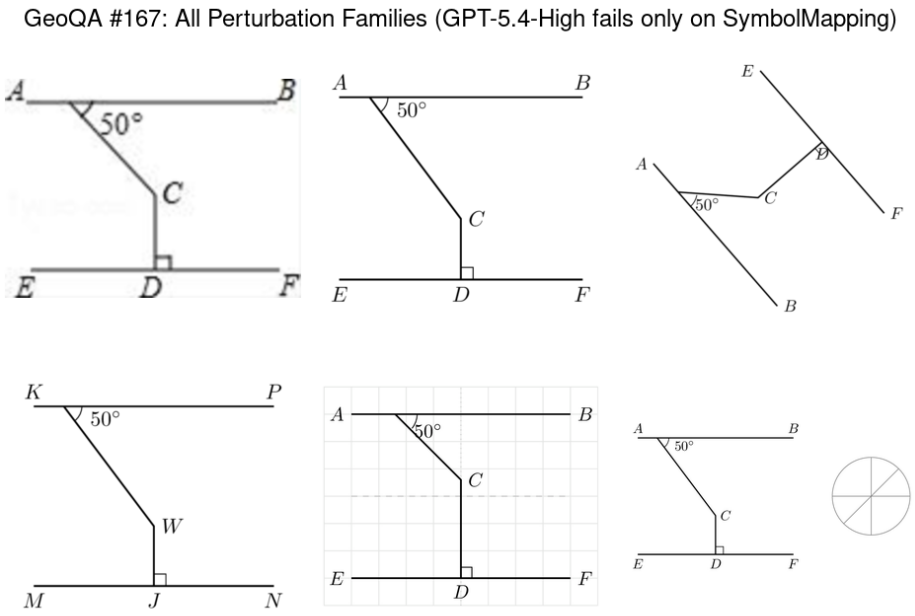

Figure 6: GeoQA #167: representative Referentiality failure under synchronized label remapping.

form 95% confidence intervals. For suite-level consistency, the permutation reference in Appendix B.1 preserves each family's marginal retention while removing cross-family dependence.

The main conclusions rely on matched effect sizes, conditional retention, and repeated patterns across model families and perturbation dimensions rather than on a single isolated cell. Since the paper contains many model–dataset cells, individual cells are interpreted descriptively unless supported by large effect sizes and consistent cross-family patterns. Because `Pure` contains 88 suites and some conditional denominators are small, Pure and low-count conditional cells are interpreted descriptively. Where multiplicity adjustments are applied, the Holm or Bonferroni procedure is reported alongside the corresponding analysis.

F Qualitative Diagnostics

We include representative cases that illustrate separable failure modes. These examples are diagnostic rather than exhaustive.

Referentiality failure. In GeoQA #167, GPT-5.4-High solves the canonical reconstruction but fails after synchronized label remapping. The geometry is unchanged; the failure is triggered by changing the symbolic identities attached to the same visual entities. This example illustrates why Referentiality is a direct test of referential grounding.

Orientability failure. In Geometry3K #2432, GPT-5.4-High solves the canonical rectangle-area problem but fails after a rigid rotation that preserves side lengths, right angles, and the gold answer. The model outputs make the transition especially concrete:

> `Canonical: Step 1: identify dimensions – length` $= 16$ `in., width` $= 9$ `in.;` $A = 16 \times 9 = 144$. [correct]
>
> `Orientability: From the diagram: length` $= 16$ `in., width` $= 6$ `in.;` $A = 16 \times 6 = 96$. [incorrect]

The model retains the correct area operation but changes the perceived width from 9 to 6 after rotation. This example shows a concrete way in which an answer-preserving redraw can break a previously correct solution without changing the underlying geometry.

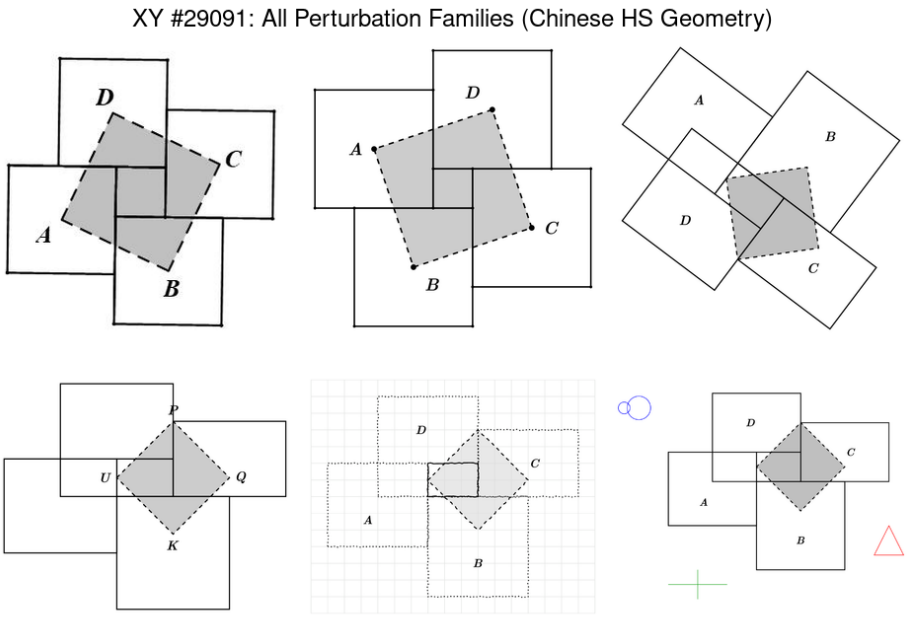

Figure 7: Pure split example #29091: algebraic geometry problem with four perturbation families.

Robust triplet. In GeoQA #0, GPT-5.4-High solves the original, canonical, and perturbed views using the same parallel-line angle relation. This case illustrates the cross-view solution stability that GEOPERTURB is designed to measure.

Pure split algebraic geometry. The Pure split also includes algebraic geometry instances rather than only angle-chasing problems. Figure 7 illustrates a tile-area problem where perturbations preserve the underlying geometric problem while exposing model-specific reasoning instability.

G PERTURBATION-GUIDED FINETUNING DETAILS

The finetuning experiment in Section 4.4 tests whether derivation supervision on perturbed views improves cross-view performance. All controlled conditions use `Qwen3-VL-8B-Instruct` and are evaluated on the same Geometry3K Source-Test items and view sets. Crucially, every row in Table 4—Base, derivation-supervised, and AnswerOnly—uses the same CoT-formatted evaluation prompt and the same scoring pipeline. The scorer evaluates final-answer equivalence and does not require a reasoning trace, so neither prompt choice nor rationale presence varies the correctness criterion across conditions. The study separates the training view (`Original`, `Canonical`, or `Perturbations`) from the supervision target (derivation plus answer, or answer only).

For each matched CoT–AnswerOnly pair, the input prompt template, image–question examples, example order, optimization settings, and random seeds are identical; only the target completion differs, with either a derivation followed by the answer or the answer alone.

For all Table 4 SFT runs, we perform full-parameter language-model tuning with the vision tower frozen, using LLaMA-Factory and DeepSpeed ZeRO-2. Training uses BF16, a sequence cutoff of 4096 tokens, an effective batch size of 64, a learning rate of 5×10^{-6} with cosine scheduling and a 0.03 warmup ratio, weight decay of 0.01, two epochs, and model/data seeds of 42. The perturbation-CoT training corpus contains 6,236 source problems and 29,978 transformed examples after excluding all evaluation IDs. Full manifests, target-generation prompts, evaluation prompts, and raw outputs are included in the release package.

At fixed perturbation training inputs and evaluation protocol, `Perturb-CoT` achieves 72.0% average perturbation accuracy versus 29.4% for `Perturb-AnswerOnly` (Table 4). It also exceeds Base (60.1%) under the same evaluation protocol. These comparisons establish the benefit of derivation-

Table 13: Paired AnswerOnly SFT results comparing Canonical TikZ and Original-raster training on matched records (%). Δ is the Canonical-TikZ minus Original-raster difference; intervals are problem-level paired 95% bootstrap intervals from 10,000 resamples, and p is the exact two-sided McNemar value.

Evaluation set	Canonical-SFT	Original-SFT	Δ	95% CI	Exact p
Geometry3K Source-Train	33.42	33.33	+0.09	[–2.26, +2.44]	1.000
Geometry3K Source-Test	30.71	29.88	+0.83	[–3.32, +4.98]	0.845
GeoQA Source-Train	72.16	70.88	+1.28	[–1.07, +3.64]	0.327
GeoQA Source-Test	64.57	63.78	+0.79	[–4.72, +6.30]	0.885
Pure	31.25	28.79	+2.46	[0.00, +4.92]	0.072

supervised training in the tested `Qwen3-VL-8B-Instruct` setup, without identifying an internal reasoning mechanism.

G.1 Rendering-Source Control

To test whether the training image format itself explains the SFT contrast, we compare two matched AnswerOnly conditions whose training records differ only in the image source: one uses Canonical TikZ renders and the other uses Original rasters. After removing empty or corrupt entries and intersecting by subset, problem ID, question, and answer, the two conditions contain 6,077 matched training records in the same order, with identical prompts and targets and only the image path changed. Both runs use `Qwen3-VL-8B-Instruct` with the vision tower frozen and otherwise matched training settings. Both rendering-source runs use three epochs; the remaining optimization settings match the controlled SFT configuration above.

Table 13 reports the paired evaluation results. Canonical-SFT is numerically higher on all five evaluation sets, but the differences are small, ranging from 0.09 to 2.46 points, and every 95% bootstrap confidence interval overlaps zero. For `Pure`, the exact McNemar test gives $p = 0.072$ (uncorrected). The rendering-source differences are small relative to the CoT–AnswerOnly contrast in Table 4.

GeoQA Source-Train is not a held-out evaluation set for this control: 55 of its problem IDs overlap the training archive, so that row is reported descriptively.

H Reproducibility Details

H.1 Rendering Configuration

Canonical and perturbation views use the same compilation and rasterization pipeline. Original-to-Canonical is therefore analyzed separately from Canonical-to-variant effects.

Table 14: TikZ compilation and rasterization configuration.

Parameter	Setting
Compiler	pdfLaTeX, TeX Live 2023
Rasterizer	ImageMagick `convert`
Resolution / background	300 DPI; white background; alpha removed
Document / padding	`standalone[tikz, border=5pt]`
Font / line width	Computer Modern by default; per-item TikZ line settings
Bounding box	Content-dependent; dimensions and padding recorded per item
Within-suite	Canonical and all variants use the same pipeline

H.2 Data Provenance and Release Policy

The release separates third-party source assets from author-created derivatives and preserves upstream terms.

Table 15: Source-specific provenance and release policy.

Resource	Provenance and release policy
GeoQA / Geometry3K	Preserve official IDs, attribution, and upstream terms. Redistribute source assets only where permitted; otherwise provide IDs, hashes, and deterministic rebuild scripts.
Pure	Release item-level source, collection/OCR/normalization status, selection criteria, duplicate-screening record, and source terms; redistribute only items whose source terms permit it.
Author-created artifacts	Release canonical TikZ, four variants, transformation code and parameters, automated flags, correction logs, final audit labels, scorer, prompts, and evaluation pipeline under compatible permissive terms.
Model evaluation	Release exact model/checkpoint identifiers, access dates, decoding settings, prompts, environment manifest, statistical scripts, and outputs where provider terms permit.

H.3 Model Identifiers and Decoding

Table 16 lists the display names and identifiers used in the evaluation. All models are evaluated with deterministic decoding, temperature 0, and a maximum output budget of 8192 tokens. For local open-source models, inference is run through the corresponding local endpoint; provider models are evaluated through their APIs. The data-generation pipeline is separate from these evaluated models: reconstruction, perturbation, and CoT-derivation generation use the gateway alias `gemini-3-flash-preview-all` (Gemini 3 Flash Preview family) through an OpenAI-compatible endpoint. Access dates, environment versions, and raw outputs are recorded in the release manifest.

Table 16: Model identifiers, providers, and evaluation configurations.

Display name	API / checkpoint identifier	Provider	Type	Datasets
Proprietary Models				
GPT-5.4-High	`gpt-5.4` with `reasoning_effort=high`	OpenAI	API	all
GPT-4o-mini	`gpt-4o-mini`	OpenAI	API	all
Claude-4.6-Opus	`claude-opus-4-6-thinking`	Anthropic	API	all
Claude-4.6-Sonnet	`claude-sonnet-4-6`	Anthropic	API	all
Gemini-3.1-Pro	`gemini-3.1-pro-preview`	Google	API	all
Qwen3-VL-Plus	`qwen3-vl-plus`	Alibaba	API	all
Open-source Models				
Qwen3-VL-8B	`Qwen/Qwen3-VL-8B-Instruct`	Alibaba / local checkpoint	local	all
InternVL3.5-8B	`OpenGVLab/InternVL3_5-8B`	Shanghai AI Lab / local checkpoint	local	all
LLaVA-OneVision-1.5-8B	`lmms-lab/LLaVA-OneVision-1.5-8B-Instruct`	LMMs-Lab / local checkpoint	local	all
MathCoder-VL-8B	`MathLLMs/MathCoder-VL-8B`	local checkpoint	local	all
G-LLaVA-13B	`renjiepi/G-LLaVA-13B`	local checkpoint	local	all
Validation-only scoring judge				
DeepSeek-V3.2	`DeepSeek-V3.2`	DeepSeek	API	validation only

H.4 Prompt Templates

Reconstruction prompt.

```
You are a Geometry-to-TikZ Transpiler. Convert the input geometry image
into precise TikZ code.  Output only the TikZ picture body.  Do not
include a document preamble, markdown, comments, or explanatory text.
Preserve geometric primitives, right-angle markers, angle arcs, labels,
point markers, and numerical annotations. Ensure labels do not overlap
whenever possible.
```

Reconstruction refinement prompt.

> Compare the original image with the rendered TikZ image. Check geometric primitives, right-angle markers, angle arcs, labels, point markers, segment labels, and spatial relations. If the render matches the original sufficiently for solving the problem, output only <PERFECT>. Otherwise, output corrected TikZ code only.

Perturbation prompt.

> You are creating a structure-preserving geometry perturbation. Target family: {family}. Preserve all lengths, angles, incidences, and answer-critical relations unless the family explicitly requires synchronized symbolic relabeling. The gold answer must remain unchanged except for Referentiality, where labels in the question, diagram, options, and symbolic answer must be updated consistently. Output only the perturbed TikZ code.

Zero-shot evaluation prompt.

> Solve the geometry problem using the provided diagram and text. Return the final answer in the following JSON format: {"answer": "<final answer>"}. The answer should be a single option, number, expression, or symbolic value whenever possible.

Shared SFT evaluation prompt. The following prompt is used verbatim for Base, derivation-supervised, and AnswerOnly conditions in Table 4. In the matched CoT–AnswerOnly comparisons, the evaluation instruction is therefore identical; the supervision target is the experimental difference.

> Solve the geometry problem using the provided diagram and text. Provide concise step-by-step geometric reasoning, then return the final answer in the following JSON format: {"answer": "<final answer>"}.

Scoring uses final-answer equivalence only. A response is not marked incorrect merely because it omits the requested reasoning trace.

Fallback judging prompt.

> You are grading a geometry answer. Given the problem, gold answer, and model response, decide whether the response contains a mathematically equivalent final answer. Ignore harmless formatting differences. Output only CORRECT or INCORRECT.

SFT target-generation (derivation) prompt. CoT training targets are produced by a derivation generator conditioned on the problem, the correct answer, and the training image:

> Produce a complete reasoning trace following these principles: (1) precise visual perception — describe what is actually in the figure; (2) robustness to visual perturbations — see through rotation, style, and distractor changes; (3) figure–text alignment — map every textual reference to a visual counterpart; (4) rigorous step-by-step derivation — name the theorem at each step and show all arithmetic. Output structure: Visual Analysis / Distractor Check / Element Mapping / Strategy / Step-by-Step Solution / Final Answer.