SEEING IN 2D, REASONING IN 3D: WHAT EXPLICIT GEOMETRY REVEALS ABOUT VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Does making geometry explicit remove the need for vision and is it enough for 3D reasoning? End-to-end accuracy alone cannot distinguish the benefit of supplying geometric information from the ability to solve problems once that information is available. We introduce GeoText, a structured representation for diagnostic input interventions over four geometric components: view elements, 3D shape structure, given measurements and constraints, and spatial relations. Using OrthoMind-Bench, a collection of 5,357 Chinese K–12 solid-geometry problems, we evaluate six open-weight vision-language models under single-component and cumulative interventions, each with and without images. Providing 3D shape descriptions yields the largest standalone gains across all six models, improving accuracy by 13.2–28.7 percentage points; substantial gains persist even after view descriptions are supplied. Additional givens provide much smaller gains once views and shape are available. Explicit geometry also changes the marginal value of vision: images improve accuracy by 10.5–27.9 points without GeoText, but their contribution becomes near-zero or negative in four models when all GeoText components are supplied. Yet, with both images and all GeoText components, model error rates remain 32.6–47.8%. These results reveal a separation between reduced visual dependence and successful geometric problem solving: external geometric descriptions can substantially reduce the benefit of images while leaving considerable failures unresolved. GeoText provides a diagnostic framework for measuring which geometric information improves performance, how it changes the value of vision, and what failures persist under explicit geometric support.

Keywords: Vision-Language Models, 3D Geometric Reasoning, Structured Geometric Representations, Diagnostic Interventions, Visual Dependence

1 INTRODUCTION

Seeing a two-dimensional diagram and reasoning about the three-dimensional object it depicts place distinct demands on a vision-language model (VLM). Lines must be interpreted as projections of spatial structure, measurements attached to the appropriate elements, and geometric constraints used to support inference. Progress in visual mathematics (Lu et al., 2024; Wang et al., 2024) and spatial reasoning (Cheng et al., 2024; Yang et al., 2025) makes understanding this transition increasingly important. Yet a wrong answer leaves a central ambiguity: did the model fail to obtain a useful geometric description, or did it fail despite having access to the relevant information?

Existing evaluations illuminate different aspects of this ambiguity. SolidGeo measures reasoning across solid-geometry tasks (Wang et al., 2025a); MathVerse varies the information distributed between text and diagrams (Zhang et al., 2024b); and We-Math and DynaSolidGeo examine knowledge composition and reasoning processes (Qiao et al., 2025; Wu et al., 2025a). These perspectives motivate a further question: *which geometric information changes the value of vision, and what remains unresolved once that information is supplied?* Removing an image alone cannot answer this question: its marginal benefit depends on what the accompanying text already makes explicit.

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.

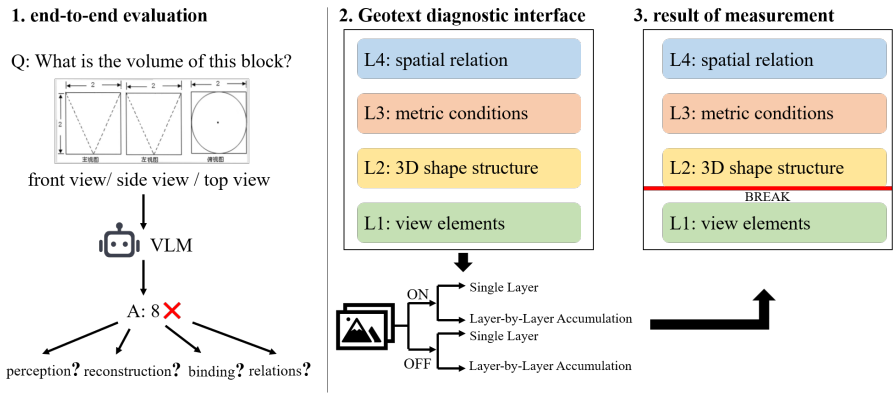


Figure 1: **GeoText diagnostic framework.** Geometric information is organized into four composable components: view elements, 3D shape structure, given measurements and constraints, and spatial relations. Single-component and cumulative interventions are evaluated with and without images to measure which descriptions improve performance, how they change the marginal value of vision, and what failures remain under explicit geometric support.

Structured representations offer a way to make this dependence measurable. Scene and question representations support diagnostic evaluation in CLEVR and GQA (Johnson et al., 2017; Hudson & Manning, 2019), while Inter-GPS and FormalGeo demonstrate the utility of explicit geometric descriptions for symbolic reasoning (Lu et al., 2021; Zhang et al., 2023). We use structured geometry as an *input intervention*: selectively supplying descriptions of the same problem and measuring both their contribution to performance and their effect on the benefit of retaining the image. This distinguishes the value of geometric support from success under that support.

We introduce **GeoText** (Figure 1), a structured representation with four composable components: view elements, 3D shape structure, given measurements and constraints, and spatial relations. On **OrthoMind-Bench**, a collection of 5,357 Chinese K–12 solid-geometry problems, we evaluate six open-weight VLMs using single-component and cumulative interventions, each with and without images. This design measures standalone gains, gains conditional on preceding components, and the marginal value of vision under matched description conditions. The components organize potentially overlapping information; they do not assume that models implement four discrete cognitive stages. Our analysis asks what information helps, when vision still adds value, and whether explicit geometric support is enough.

Three findings emerge. First, shape descriptions provide the largest standalone gains across all six models, improving accuracy by 13.2–28.7 percentage points; substantial gains remain after view descriptions are supplied. Second, explicit geometry changes the value of vision: images add 10.5–27.9 points without GeoText, but their contribution becomes near-zero or negative in four models when all available components are supplied. Third, even with images and all components, error rates remain 32.6–47.8%. Together, these results expose a separation between reduced visual dependence and reliable geometric reasoning. External descriptions can substantially reduce the benefit of images while leaving considerable failures unresolved.

Our contributions are a composable geometric intervention framework, an annotated evaluation resource, and an empirical characterization of how geometric support changes both visual dependence and residual failure. GeoText makes these outcomes separately measurable, providing a more informative account of what explicit geometry enables and what it leaves unresolved.

2 GeoText: A Framework for Diagnostic Interventions

GeoText treats the geometric information available to a model as an experimental variable. We vary the supplied descriptions and image availability while retaining the original question. The resulting comparisons measure which descriptions improve performance, how their value depends on other descriptions, and whether images remain useful under the same geometric support. The framework

Table 1: **GeoText components and non-empty coverage.** Coverage is measured over all 5,357 problems and does not establish correctness or completeness. Shared content makes component gains potentially non-additive. Batch-level coverage is reported in Appendix C.

Component	Supplied information	Shared content	Non-empty coverage
Views (V)	Projection outlines, internal lines, and dimension labels	Dimensions also occur in Shape and Givens	86.4%
Shape (S)	Solid type, composition, arrangement, and dimensions	Sizes and relationships overlap with Givens and Relations	99.5%
Givens (G)	Stated measurements and constraints bound to elements	Values and constraints also occur in other components	90.5%
Relations (R)	Spatial constraints and relationships between elements	Relationships can also occur in Shape or Givens	33.4%

evaluates responses to external information rather than presupposing an internal sequence of cognitive stages.

2.1 Geometric Information Components

Each description comprises four selectable components, $\mathcal{C} = \{\mathrm{V}, \mathrm{S}, \mathrm{G}, \mathrm{R}\}$. **Views** records projection outlines, internal lines, and dimension labels. **Shape** describes the three-dimensional object, including its type, composition, arrangement, and associated dimensions. **Givens** binds stated measurements and constraints to geometric elements. **Relations** describes spatial relationships, including parallelism, perpendicularity, projection, and inscription. Table 1 summarizes their content and availability; Appendix A provides the schema and a worked example.

The components distinguish forms of description without requiring semantic independence. For example, a solid's height may appear both in Shape and as a labeled dimension in Views. Their overlap is informative: comparing standalone and conditional gains reveals whether a description remains useful after other information is supplied. These gains characterize the utility of the provided descriptions; they do not isolate purely structural, perceptual, or numerical capabilities.

2.2 Benchmark and Description Construction

OrthoMind-Bench contains 5,357 Chinese K–12 solid-geometry problems with question text, figures, and reference answers. Three-view drawings constitute 80.2% of the collection; the remainder includes solid diagrams and nets. Open-form answers account for 75.1%, and multiple-choice answers for 24.9%. Two batches of 1,420 and 3,937 problems have disjoint identifiers and are not difficulty strata. Thus, the aggregate evaluation primarily concerns diagram-based solid geometry, rather than exclusively three-view reconstruction. Appendix C reports the composition and automatic labeling protocol.

GeoText is generated with Gemini-3.1-Pro-Preview and stored as JSON with Chinese natural-language glosses. Selected components are rendered as labeled blocks, separately for each solid when needed. Only `metrics.given` is supplied; `metrics.derived`, which can contain computed quantities or answers, is excluded. This removes the explicitly stored derived branch, but does not guarantee that retained descriptions never overlap with a task's target. Appendix A makes this distinction concrete through a volume example.

Coverage materially affects interpretation. Shape is non-empty for 99.5% of problems, whereas Relations is non-empty for 33.4%; the latter is 100% in the 1,420-problem batch but only 9.4% in the 3,937-problem batch. A small aggregate Relations gain therefore cannot establish that spatial relations are unimportant. Moreover, empty dictionaries remain labeled blocks, while missing keys

suppress a block, so empty-content interventions need not reproduce the baseline prompt. Throughout, *full GeoText* means all available components, not complete or certified geometric information. Performance under this support is an observed outcome, not an oracle upper bound.

2.3 Intervention Design and Measurements

Let $A_m^{(q)}(U, z)$ be model m's accuracy with component subset $U \subseteq \mathcal{C}$ and image indicator $z \in \{0, 1\}$, in intervention series q. The *single-component* series supplies one component at a time. The *cumulative* series follows $(c_1, c_2, c_3, c_4) = (\mathrm{V}, \mathrm{S}, \mathrm{G}, \mathrm{R})$, with $U_0 = \varnothing$ and $U_k = \{c_1, \ldots, c_k\}$. Their shared baseline and Views-only information sets yield eight distinct component subsets, or sixteen information configurations with images present or absent. We retain the series index because shared information sets need not have identical introductory wording. Appendix B gives the condition mapping and prompt templates.

Within each series, the original question is retained and image-present conditions reuse the same image. Removing the image also removes references to it from the prompt. We define:

$$\Delta_m^{\mathrm{single}}(c; z) = A_m^{(\mathrm{single})}(\{c\}, z) - A_m^{(\mathrm{single})}(\varnothing, z), \tag{1}$$

$$\Delta_m^{\mathrm{cond}}(c_k; z) = A_m^{(\mathrm{cum})}(U_k, z) - A_m^{(\mathrm{cum})}(U_{k-1}, z), \tag{2}$$

$$\delta_m^{\mathrm{image}}(U; q) = A_m^{(q)}(U, 1) - A_m^{(q)}(U, 0). \tag{3}$$

These contrasts measure standalone benefit, incremental benefit under the fixed order, and the conditional benefit of retaining the image, respectively. Contrasts are reported in percentage points. The residual error rate under full support is $1 - A_m^{(\mathrm{cum})}(\mathcal{C}, 1)$.

A large standalone gain can diminish when overlapping descriptions are already present; a small image contrast can coexist with low accuracy. These are precisely the distinctions the design exposes. Conditional gains remain order-specific, and near-zero average image benefit does not imply that every problem is unaffected. The design does not estimate all component interactions or assign residual errors to separate internal mechanisms. Appendix B.4 details these interpretations, including why net accuracy gains are not counts of repaired errors.

2.4 Models and Evaluation

We evaluate Qwen3.5-2B, -4B, -9B, and -27B, together with GLM-4.5V and InternVL3.5-8B. The Qwen series supports within-family scale comparisons; the remaining models extend the evaluation across families. Each problem–condition pair is run once with temperature 0, top-p 0.9, a 10,000-token output limit, and thinking disabled in the evaluation configuration. The conclusions are conditional on these models and settings.

Prompts request the final answer inside `<answer>` tags. Tagged content is extracted when available; otherwise the full response is retained. Rule-based scoring uses exact and normalized string matching, with Qwen3-30B-A3B used for answer-level rechecking. This evaluates final-answer correctness, not intermediate reasoning validity. Consequently, residual errors remain unclassified: they cannot be identified as purely computational failures from these measurements alone. Appendix D documents extraction, normalization, and execution settings.

3 What Explicit Geometry Reveals

GeoText separates three questions that end-to-end accuracy leaves entangled: which geometric descriptions help, whether images remain useful once those descriptions are available, and how reliably models solve the resulting problems. The results reveal a consistent distinction between access to useful geometric information and successful problem solving under that support. Unless stated otherwise, all comparisons use the full set of 5,357 problems. Accuracy differences are reported in percentage points (pp); Appendix E provides the numerical results underlying the analyses.

3.1 WHICH GEOMETRIC INFORMATION HELPS?

Shape is the most effective standalone description. With images present, Shape yields the largest standalone gain in every model: 13.15–28.74 pp, compared with 3.98–11.61 pp for Views and 8.56–19.37 pp for Givens (Figure 2). This ranking holds across all four Qwen sizes and both other model families, despite substantial differences in their baseline accuracies. InternVL3.5-8B exhibits the largest Shape gain, rising from 24.11% to approximately 52.85% accuracy with Shape alone. The agreement across models identifies a consistently useful form of external support: a description of the solid, rather than only descriptions of its projected elements or given measurements.

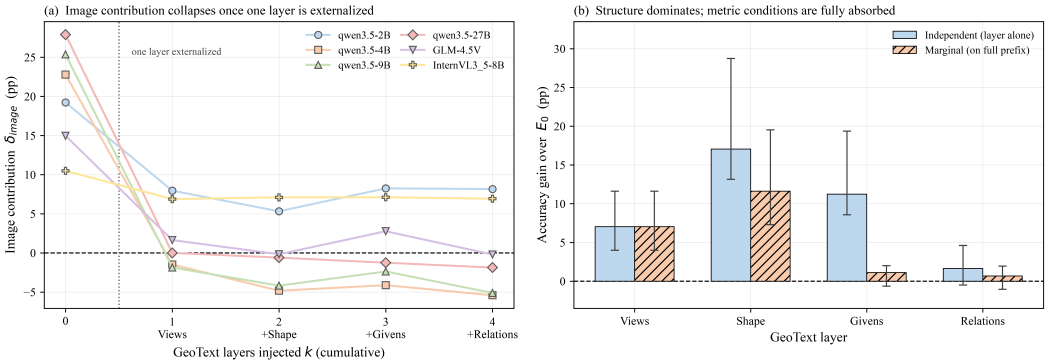


Figure 2: **Shape helps both alone and after Views.** Standalone and conditional component gains with images present, in percentage points. Standalone gains compare each component with the no-GeoText baseline; conditional gains follow Views → Shape → Givens → Relations. The two comparisons expose information overlap and order dependence; their entries are not contributions of mutually exclusive capabilities. Exact values appear in Appendix E.2.

Explicit views leave substantial additional value for Shape. After Views is supplied, adding Shape still improves accuracy by 7.29–19.52 pp. Qwen3.5-27B gains a further 7.29 pp, and InternVL3.5-8B gains 19.52 pp. Thus, making projection outlines, internal lines, and labels explicit does not exhaust the benefit of describing the solid. Standalone and conditional comparisons jointly strengthen this finding: Shape helps both when supplied alone and when a textual account of the views is already available. Because Shape includes dimensions and relationships as well as structure, this is a result about the supplied description as a whole, rather than an isolated measure of an internal reconstruction stage.

Stronger baseline performance does not eliminate this benefit. Within Qwen3.5, image-present baseline accuracy increases from 31.44% at 2B to 50.63% at 27B, while standalone Shape gains decrease from 16.72 to 13.34 pp. The improvement in baseline performance therefore accompanies a smaller, but still substantial, benefit from explicit shape information. Even the strongest baseline retains both a 13.34 pp standalone gain and a 7.29 pp gain after Views. These observations establish persistence over the evaluated scale range, without extrapolating to an asymptotic scaling law.

Standalone utility and incremental utility differ sharply. Givens has substantial standalone value, yet adds only -0.64 to $+2.00$ pp after Views and Shape. For InternVL3.5-8B, its gain falls from 19.37 pp alone to 0.87 pp conditionally. This reduction is consistent with measurements already appearing in the preceding descriptions; it does not show that numerical binding is solved. Nor does Shape dominate every cumulative step: the Views increments for Qwen3.5-9B and -27B are slightly larger than their subsequent Shape increments. The consistent finding is that Shape is the best standalone component and remains useful after Views, not that one component explains every improvement.

Small Relations gains require a coverage-aware interpretation. Relations contributes -0.50 to $+4.60$ pp alone and -1.05 to $+1.94$ pp at the final cumulative step. Its non-empty coverage is only 33.4% overall (Table 1). For Qwen3.5-27B, however, the conditional gain remains small in both

batches: 0.74 pp with 100% Relations coverage and 0.64 pp with 9.4% coverage (Appendix E.4). Missing content alone therefore cannot explain this model's small increment. The comparison remains conditional on preceding descriptions that may already express relations, and the batches differ in composition. It neither establishes that relations are unnecessary nor extends the batch-level observation to all six models.

3.2 When Does Vision Still Add Value?

Image utility changes when geometry becomes explicit. Without GeoText, images improve accuracy in all six models, by 10.49–27.89 pp. With full GeoText, this contrast decreases in every model (Figure 3). It becomes -5.41, -5.12, and -1.87 pp for Qwen3.5-4B, -9B, and -27B, respectively, and -0.20 pp for GLM-4.5V. For Qwen3.5-27B, the same image modality that contributes 27.89 pp at baseline has a negative aggregate contribution under full support. Baseline image benefit therefore does not predict an enduring need for images once geometric descriptions are supplied.

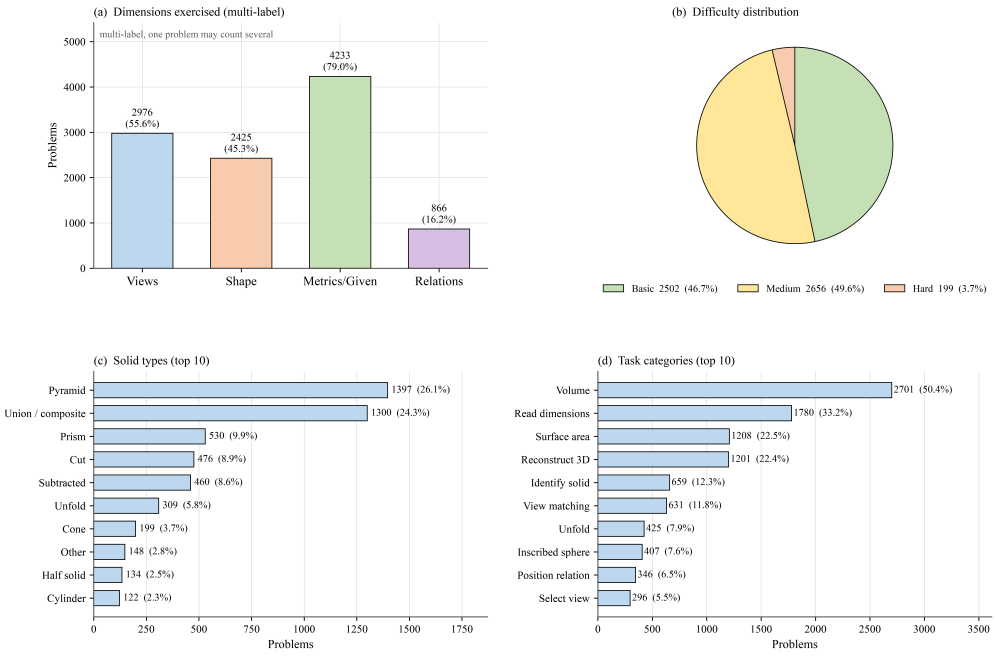


Figure 3: **Explicit geometry changes the marginal value of vision.** Image-present minus image-absent accuracy without GeoText and with all available components, in percentage points. The contrast decreases in every model; under full support it is near zero or negative in four models and remains positive in two. Connections compare the two endpoints and do not establish a monotonic trajectory through intermediate interventions. Numerical endpoints appear in Appendix E.3.

Substitution is substantial, but model-dependent. Qwen3.5-2B and InternVL3.5-8B retain image gains of 8.16 and 6.93 pp under full support. The supplied descriptions do not exhaust the useful contribution of the image for these models. Conversely, the negative contrasts in larger Qwen models show that retaining previously helpful visual input can reduce aggregate accuracy once explicit geometry is available. Together, the common reduction and the different endpoints indicate that visual dependence is conditional on both the model and the information supplied alongside the image.

GeoText helps more when the image is absent. Full GeoText improves text-only accuracy by 31.78–46.91 pp, exceeding the corresponding image-present gain in every model. For Qwen3.5-9B, text-only accuracy rises from 22.46% to 69.37%, while image-present accuracy rises from 47.82% to 64.25%. These unequal gains account for the reduction in the image contrast: descriptions improve performance in both configurations, but improve the text-only configuration more. Thus, the ob-

served reduction in visual dependence accompanies substantial improvements in task performance, rather than merely convergence at a low accuracy.

The endpoint comparison identifies changing utility. This experiment holds the question and available GeoText fixed when comparing image-present and image-absent conditions. It therefore asks whether the image adds value under that support, rather than whether the task was originally visual. The small GLM contrast is an observed difference, not a test of equivalence; negative averages likewise do not prescribe image removal for every problem. These results establish a change between the baseline and full-support endpoints, without requiring image benefit to decline monotonically at intermediate steps. The evidence concerns the effect of retaining an image, rather than whether or how the model internally attends to it.

3.3 Is Explicit Geometry Enough?

Substantial improvements leave substantial failure. With images and all available GeoText components, accuracy improves over baseline by 16.43–33.05 pp, but reaches only 52.16–67.37%. Thus, 32.63–47.84% of problems remain incorrect. Figure 4 places these residual errors alongside the corresponding image contrasts. The four models with near-zero or negative image benefit still have error rates of 32.63–46.18%. Qwen3.5-9B, for example, loses 5.12 pp when the image is retained, yet answers 35.75% of problems incorrectly with it. The two models that continue to benefit from images also retain high error rates: 47.84% for Qwen3.5-2B and 42.84% for InternVL3.5-8B.

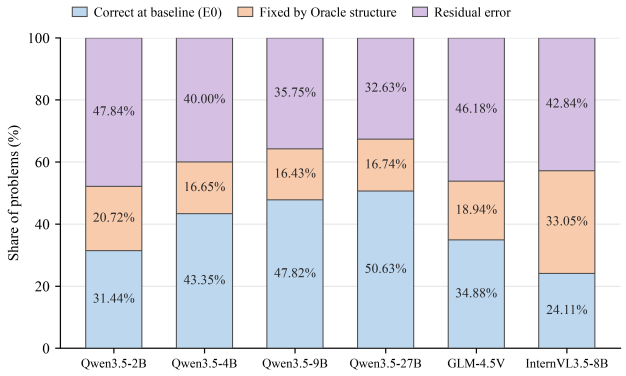


Figure 4: **Low image utility does not imply reliable problem solving.** Each model is positioned by its image contrast under full GeoText (horizontal axis, pp) and its error rate with images and full GeoText (vertical axis, %). Substantial error remains on both sides of zero image benefit. The coordinates describe outcomes under the supplied support, not a decomposition into visual and reasoning errors. Exact coordinates appear in Appendix E.3.

Removing images does not close the remaining gap. These failures are not explained simply by retaining an image whose average contribution has become negative. Without images, full-GeoText accuracy reaches at most 69.37% (Table 2); even the best observed text-only result leaves 30.63% incorrect. Both input configurations therefore exhibit the same separation: making geometric descriptions available can sharply reduce the additional value of vision while leaving considerable failures unresolved. An image contrast near zero measures little net benefit from retaining the image, not proximity to a correct solution.

Full support is an evaluated endpoint, not a performance ceiling. Adding a component need not improve accuracy even within the image-present cumulative sequence. Qwen3.5-2B decreases from 51.56% to 50.92% when Givens follows Views and Shape; Qwen3.5-9B decreases from 65.30% to 64.25% when Relations is added last (Appendix E.1). Full GeoText is therefore neither a guaranteed best input configuration nor certified complete geometry. Its remaining errors measure performance under the supplied support, rather than a model's maximum attainable accuracy or a purely computational error category.

Table 2: **Key accuracy comparisons on 5,357 problems (%).** Every condition retains the original question. I and T denote image-present and image-absent baselines; S denotes Shape alone and All denotes all available components. $I{+}S$ values are reconstructed from the reported baseline and standalone Shape gain, both rounded to two decimals. Other entries are reported condition accuracies.

Model	I	$I{+}S$	$I{+}$All	T	$T{+}$All
Qwen3.5-2B	31.44	48.16	52.16	12.22	44.00
Qwen3.5-4B	43.35	59.29	60.00	20.56	65.41
Qwen3.5-9B	47.82	62.24	64.25	22.46	69.37
Qwen3.5-27B	50.63	63.97	67.37	22.74	69.24
GLM-4.5V	34.88	48.03	53.82	19.91	54.02
InternVL3.5-8B	24.11	52.85	57.16	13.62	50.23

Net improvement does not identify which errors were repaired. The 16.43–33.05 pp gains count corrected answers minus newly incorrect answers. Aggregate accuracy does not establish how many baseline failures were repaired or how many full-support failures were introduced by the intervention. Appendix B.4 formalizes this distinction. The supported conclusion is consequently a separation between observable outcomes: explicit shape descriptions are consistently useful, images can lose much of their additional value, and reliable geometric problem solving remains unresolved under either modality configuration.

4 Implications and Limitations

Visual dependence is not a certificate of spatial understanding. The central finding is a separation between the marginal utility of an observation and competence on the task it describes. Once geometry is explicit, an image can contribute little additional accuracy even while substantial failure remains. Low visual dependence therefore cannot establish that a model has acquired a sufficient spatial understanding; it describes what retaining the image adds under a particular information configuration. Conversely, the usefulness of images at baseline does not establish that their geometric content is represented in a form the model can reliably use. Evaluations of spatial intelligence should consequently report information utility and supported task performance together.

An intermediate representation must be both faithful and usable. Shape's persistent benefit after Views motivates investigating the interface between visual observations and usable spatial descriptions. The question is not simply whether geometric information is present, but whether its organization enables the model to exploit it. This suggests evaluating spatial representations by fidelity to the depicted geometry and utility for downstream decisions. A faithful description may leave reasoning difficult, while a helpful description may encode task-specific information that does not generalize. GeoText establishes the utility of supplied descriptions, without establishing their complete fidelity or autonomous recoverability. For systems generating their own spatial representations, generation quality, downstream utility, and end-to-end performance must be assessed separately. Improvements with externally supplied geometry motivate such interfaces; they do not demonstrate that a self-generated interface preserves those improvements.

Embodied models require geometry grounded in action. This perspective extends to VLMs used as robotic reasoning modules and vision–language–action (VLA) models predicting actions from visual and linguistic inputs. Their spatial descriptions could expose object shape, pose, relative placement, and contact or support relations, allowing interventions on three-dimensional information available before decisions. The extension is to evaluate whether this information supports physical action. A description sufficient for object identification need not specify the scale, reference frame, reachability, or clearance required to manipulate it. A robotic counterpart to GeoText could compare scene interpretation and action outcomes under supplied spatial information, while separately assessing its recoverability from observations. These comparisons could test whether better spatial descriptions improve decisions. They would not identify unique internal stages, and reduced average image benefit would not justify discarding visual feedback: motion, contact, and uncertainty can make previously supplied descriptions insufficient.

Scope and limitations. GeoText measures the conditional utility of geometric descriptions: component overlap, intervention order, and description quality shape the observed gains. Our findings concern six open-weight VLMs on Chinese K–12 geometry under a fixed inference configuration; their broader applicability remains to be explored. Small accuracy differences warrant cautious interpretation, and understanding residual failures requires evidence beyond final answers.

5 Related Work

Visual mathematics and diagnostic evaluation. Multimodal benchmarks assess broad reasoning (Yue et al., 2024; Chen et al., 2024b; Lu et al., 2024; Wang et al., 2024), while geometry benchmarks target calculation and proof (Chen et al., 2021; 2022; Zhang et al., 2024a). MathVerse varies visual–textual information (Zhang et al., 2024b), and We-Math decomposes knowledge-dependent reasoning (Qiao et al., 2025). Specialized training improves mathematical understanding (Gao et al., 2023; Xia et al., 2025; Zhang et al., 2025b; Shi et al., 2024). These studies motivate finer diagnosis beyond aggregate accuracy. GeoText crosses geometric-component interventions with image availability. By distinguishing view elements, 3D shape structure, givens, and spatial relations, it tests which information helps, both independently and conditional on other components, on the same problems without retraining.

Spatial understanding. Diagnostic visual reasoning (Johnson et al., 2017; Hudson & Manning, 2019), perception and spatial benchmarks (Fu et al., 2024; Liu et al., 2023; Du et al., 2024; Yang et al., 2025; Wang et al., 2025b), and solid-geometry evaluation (Wang et al., 2025a) probe complementary abilities; DynaSolidGeo additionally evaluates reasoning processes (Wu et al., 2025a). Spatial supervision, depth, and geometric features improve spatial models (Chen et al., 2024a; Cheng et al., 2024; Cai et al., 2024; Zhang et al., 2025a; Wu et al., 2025b; Song et al., 2025). Our focus is how supplied descriptions change image utility and what failures persist. This distinction matters because improvements from additional geometric information and the ability to reason with that information are separate empirical questions. The interventions characterize observable input dependence without presupposing a particular internal representation.

Structured representations and embodied models. Formal geometry solvers (Lu et al., 2021; Zhang et al., 2023), 3D question answering and language models (Azuma et al., 2022; Ma et al., 2023; Hong et al., 2023; Huang et al., 2024), and embodied models (Driess et al., 2023; Zitkovich et al., 2023; Kim et al., 2025) connect representations to answers or actions. Affordance prediction, relational constraints, and embodied reasoning further support robotic decisions (Yuan et al., 2025; Huang et al., 2025; Zawalski et al., 2025). GeoText uses structured geometry as a diagnostic input interface, jointly measuring component utility, conditional image benefit, and residual error. It complements representation-learning and solving approaches without equating reduced visual dependence with successful geometric reasoning. These connections motivate broader applications of the diagnostic principle; the present evidence concerns solid-geometry problems.

6 Conclusion

We introduced GeoText to investigate what explicit geometric information contributes to VLMs' reasoning from 2D diagrams about 3D objects. By varying four geometric components with and without images, our evaluation separates the benefit of supplied information, the remaining contribution of vision, and performance under explicit geometric support. Across six models, descriptions of 3D shape structure provide the largest standalone gains and remain valuable after view descriptions are available. With all components supplied, the marginal benefit of images becomes near-zero or negative in four models, yet error rates with both images and GeoText remain 32.6–47.8%. Thus, reduced visual dependence and successful geometric reasoning are distinct outcomes: making geometry explicit can substantially change what images contribute without resolving the task. GeoText provides a diagnostic framework for studying this distinction. More broadly, evaluating spatial intelligence requires asking both what information a model needs from visual input and what it can accomplish once that information is available.

AI USE STATEMENT

Generative AI tools were used only for literature exploration, reference checking, and language polishing of selected sentences in the final version of the manuscript. The research ideas, overall logic, methodological design and description, and experimental analysis were developed independently by the authors. All technical content, citations, equations, implementations, experimental results, and scientific claims were manually reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Section 2 defines the GeoText representation and the single-component and cumulative intervention protocols, including the corresponding conditions with and without images. Section 3 presents the comparisons across six open-weight VLMs, with detailed numerical results provided in Appendix E. Evaluating the same problems across input conditions supports direct comparison of component utility, conditional image benefit, and residual error. Upon acceptance, we will release the GeoText annotations, intervention-construction code, model prompts and evaluation configurations, and scoring and analysis scripts to facilitate independent reproduction and further investigation.

REFERENCES

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. ScanQA: 3D question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19129–19139, 2022.

Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. SpatialBot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. *arXiv preprint arXiv:2401.12168*, 2024a.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P. 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. doi: 10.18653/v1/2021.findings-acl.46.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. UniGeo: Unifying geometry logical reasoning via reformulating mathematical expression. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3313–3323, 2022. doi: 10.18653/v1/2022.emnlp-main.218.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, 2024b.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. EmbSpatial-Bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 346–355, 2024. doi: 10.18653/v1/2024.acl-short.33.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, 2024.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. G-LLaVA: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3D-LLM: Injecting the 3D world into large language models. *arXiv preprint arXiv:2307.12981*, 2023.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3D world. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20413–20451, 2024.

Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. ReKep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 4573–4602, 2025.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Fei-Fei Li, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713, 2025.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. doi: 10.1162/tacl_a_00566.

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. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.528.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: Situated question answering in 3D scenes. In *International Conference on Learning Representations*, 2023.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma GongQue, Shanglin Lei, YiFan Zhang, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Xiao Zong, Yida Xu, Peiqing Yang, Zhimin Bao, Muxi Diao, Chen Li, and Honggang Zhang. We-Math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20023–20070. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.983.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-LLaVA: Bootstrapping mathematical reasoning for multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4663–4680, 2024. doi: 10.18653/v1/2024.findings-emnlp.268.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. RoboSpatial: Teaching spatial understanding to 2D and 3D vision-language models for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15768–15780, 2025.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-Vision dataset. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Peijie Wang, Chao Yang, Zhong-Zhi Li, Fei Yin, Dekang Ran, Mi Tian, Zhilong Ji, Jinfeng Bai, and Cheng-Lin Liu. SolidGeo: Measuring multimodal spatial math reasoning in solid geometry. In *Advances in Neural Information Processing Systems*, volume 38, 2025a.

Siting Wang, Luoyang Sun, Cheng Deng, Kun Shao, Minnan Pei, Zheng Tian, Haifeng Zhang, and Jun Wang. SpatialViz-Bench: Automatically generated spatial visualization reasoning tasks for MLLMs. *arXiv preprint arXiv:2507.07610*, 2025b.

Changti Wu, Shijie Lian, Zihao Liu, Lei Zhang, Laurence Tianruo Yang, and Kai Chen. Dyna-SolidGeo: A dynamic benchmark for genuine spatial mathematical reasoning of VLMs in solid geometry. *arXiv preprint arXiv:2510.22340*, 2025a.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. doi: 10.52202/085713-0457.

Renqiu Xia, Mingsheng Li, Hancheng Ye, Wenjie Wu, Hongbin Zhou, Jiakang Yuan, Tianshuo Peng, Xinyu Cai, Xiangchao Yan, Bin Wang, Conghui He, Botian Shi, Tao Chen, Junchi Yan, and Bo Zhang. GeoX: Geometric problem solving through unified formalized vision-language pre-training. In *International Conference on Learning Representations*, 2025.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10632–10643, 2025.

Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. RoboPoint: A vision-language model for spatial affordance prediction in robotics. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 4005–4020, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 3157–3181, 2025.

Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. From flatland to space: Teaching vision-language models to perceive and reason in 3D. *arXiv preprint arXiv:2503.22976*, 2025a.

Jiaxin Zhang, Zhong-Zhi Li, Ming-Liang Zhang, Fei Yin, Cheng-Lin Liu, and Yashar Moshfeghi. GeoEval: Benchmark for evaluating LLMs and multi-modal models on geometry problem-solving. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1258–1276, 2024a. doi: 10.18653/v1/2024.findings-acl.73.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems? In *Computer Vision – ECCV 2024*, volume 15066 of *Lecture Notes in Computer Science*, pp. 169–186. Springer Nature Switzerland, 2024b. doi: 10.1007/978-3-031-73242-3_10.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Shanghang Zhang, Peng Gao, and Hongsheng Li. MAVIS: Mathematical visual instruction tuning with an automatic data engine. In *International Conference on Learning Representations*, 2025b.

Xiaokai Zhang, Na Zhu, Yiming He, Jia Zou, Qike Huang, Xiaoxiao Jin, Yanjun Guo, Chenyang Mao, Yang Li, Zhe Zhu, Dengfeng Yue, Fangzhen Zhu, Yifan Wang, Yiwen Huang, Runan Wang, Cheng Qin, Zhenbing Zeng, Shaorong Xie, Xiangfeng Luo, and Tuo Leng. FormalGeo: An extensible formalized framework for olympiad geometric problem solving. *arXiv preprint arXiv:2310.18021*, 2023.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183, 2023.

A GeoText Schema and Examples

A.1 Representation and component selection

GeoText stores geometric descriptions as JSON. Its four selectable components are `views`, `shape`, `metrics.given`, and `relations`, corresponding to V, S, G, and R in Section 2.1. Chinese natural-language glosses are stored under `cn`. Table 3 specifies the representation. These fields organize information for input interventions; they are not a partition into mutually exclusive facts or an executable geometric specification.

Table 3: **Storage schema and intervention payloads.** The stored `metrics` object can contain derived quantities, but only its `given` branch is supplied.

Field	Structure and interpretation
`views`	Projection descriptions under `front`, `side`, and `top`, with `outline`, `labeled_dims`, `inner_lines`, and `cn`; `layout` records the arrangement.
`shape`	Solid type, dimensions, and composition. A difference can use `type: Subtract` with `base` and `subtracted`; a union can use `type: Union` with `parts`. Components can themselves be solids.
`metrics`	`given` stores measurements and constraints associated with elements. `derived` stores computed quantities and is excluded from interventions.
`relations`	Spatial constraints, for example an `inscribed` list whose entries identify `inscribed_type`, `inner`, and `outer`.
`cn`	A natural-language gloss attached to an object. Glosses within selected fields are part of the supplied description.

For multiple solids, the representation is a list of objects. Selected fields are rendered separately for each object, with the solid index included in the block label. Selection operates on fields, rather than retrieving individual facts in response to the question. In particular, selecting Shape can supply both structure and dimensions, while selecting Views can supply dimensions visible in projections. Consequently, an observed Shape gain measures the utility of the complete Shape description.

A.2 Worked example: a cube with a conical cavity

The example is problem `a051b2f3bf55772579cd43d6a15255c7` from the 1,420-problem batch. Its question translates as follows:

> A special building block is obtained by removing a cone from the middle of a cube. Its corresponding three-view drawing is shown. What is the volume of the block?

The reference answer is $8 - 2\pi/3$. The following is an English rendering of the stored record: Chinese glosses and descriptive keys are translated, while the field hierarchy and numerical content are preserved. It illustrates the representation, rather than reproducing the Chinese input bytes.

This record makes three distinctions explicit. First, the question already identifies a cube with a cone removed, so the example is not evidence that the model must infer this composition from projections alone. Second, dimensions and inscription information recur across components: standalone gains cannot be interpreted as contributions of disjoint facts. Third, `metrics.derived.volume` is the reference answer itself. Excluding this branch is therefore essential. Under a Givens intervention, only the `given` branch is supplied; the displayed derived branch documents what is removed.

Branch exclusion is a concrete implementation rule, not a certification of every retained fact. In a different task, identifying the solid or a spatial relation could itself answer the question. Thus, GeoText interventions measure the effect of making descriptions explicit, including any task-relevant information they convey.

```
{
  "cn": "A cube with an inverted cone removed from its middle",
  "views": {
    "layout": "combined",
    "front": {
      "outline": {"shape": "square", "side": 2},
      "labeled_dims": {"width": 2, "height": 2},
      "inner_lines": [
        {"line_type": "dashed", "geometry": "line",
         "cn": "From the upper-left corner to the bottom midpoint"},
        {"line_type": "dashed", "geometry": "line",
         "cn": "From the upper-right corner to the bottom midpoint"}
      ],
      "cn": "A square of side 2 containing a dashed inverted isosceles
    triangle"
    },
    "side": {
      "outline": {"shape": "square", "side": 2},
      "labeled_dims": {"width": 2},
      "inner_lines": [
        {"line_type": "dashed", "geometry": "line",
         "cn": "From the upper-left corner to the bottom midpoint"},
        {"line_type": "dashed", "geometry": "line",
         "cn": "From the upper-right corner to the bottom midpoint"}
      ],
      "cn": "A square of side 2 containing a dashed inverted isosceles
    triangle"
    },
    "top": {
      "outline": {"shape": "square", "side": 2},
      "labeled_dims": {"width": 2},
      "inner_lines": [
        {"line_type": "solid", "geometry": "circle",
         "cn": "An inscribed circle with a solid dot at its center"}
      ],
      "cn": "A square of side 2 containing its inscribed circle"
    }
  },
  "shape": {
    "type": "Subtract",
    "base": {
      "type": "Prism", "base": {"shape": "square", "side": 2},
      "h": 2, "subtype": "regular", "cn": "A cube of side 2"
    },
    "subtracted": {
      "type": "Cone", "r": 1, "h": 2,
      "cn": "A cone of base radius 1 and height 2"
    },
    "cn": "A cube minus an inverted cone of equal height whose base is
    inscribed in the cube's top face"
  },
  "metrics": {
    "given": {
      "cube_edge": 2, "cone_base_diameter": 2, "cone_height": 2,
      "cn": "Cube edge 2; removed cone base diameter 2 and height 2"
    },
    "derived": {
      "volume": "8 - 2*pi/3",
      "cn": "Cube volume (8) minus cone volume (2*pi/3)"
    }
  },
  "relations": {
    "inscribed": [
      {"inscribed_type": "circle_in_polygon",
```

```
        "inner": "cone base", "outer": "cube top face",
        "cn": "The cone base is inscribed in the cube's top face"}
    ]
  }
}
```

B Intervention Protocol and Interpretation

B.1 Condition mapping

Table 4 maps the intervention series to supplied components. The single-component series consists of a baseline and four standalone interventions. The cumulative series consists of a baseline followed by Views, Shape, Givens, and Relations. Both series are evaluated with and without images.

Table 4: **Condition mapping.** Checkmarks indicate inclusion in a series. Each row is evaluated with images present and absent. Shared component subsets do not imply identical prompts or a shared execution.

Supplied components U	Single-component series	Cumulative series
$\varnothing$	✓	U_0
$\{\mathrm{V}\}$	✓	U_1
$\{\mathrm{S}\}$	✓	—
$\{\mathrm{G}\}$	✓	—
$\{\mathrm{R}\}$	✓	—
$\{\mathrm{V},\mathrm{S}\}$	—	U_2
$\{\mathrm{V},\mathrm{S},\mathrm{G}\}$	—	U_3
$\{\mathrm{V},\mathrm{S},\mathrm{G},\mathrm{R}\}$	—	U_4

There are eight distinct information subsets and sixteen subset–image combinations. This count describes information configurations, not necessarily the number of separately executed prompts. In particular, the single-component and cumulative templates use different introductory phrases. We therefore retain the series index in $A_m^{(q)}(U, z)$ and form contrasts within a series. This design is not the full factorial over four components, which would require all sixteen subsets before varying image availability.

B.2 System and user prompts

Experiments use Chinese prompts. The templates below are English translations preserving their structure; placeholders in braces are populated with the question and serialized descriptions.

System prompt, image present.

> You are a mathematics assistant skilled at solid geometry. Carefully analyze the problem and the image, and give the final answer. Place your final answer between `<answer>` and `</answer>` tags.

System prompt, image absent.

> You are a mathematics assistant skilled at solid geometry. Carefully analyze the problem and give the final answer. Place your final answer between `<answer>` and `</answer>` tags.

Baseline user prompt, image present.

> Please solve this solid geometry problem based on the problem statement and the image, and give the final answer.
> Problem: `{question}`

Cumulative user prompt, image present.

> Please solve this solid geometry problem based on the problem statement, the image, and the following supplementary given information, and give the final answer.
> Problem: `{question}`
> Among the following supplementary information, `{layer descriptions}`:
> `{labeled JSON blocks}`

Single-component user prompt, image present.

> Please solve this solid geometry problem based on the problem statement, the image, and the following supplementary given information, and give the final answer.
> Problem: `{question}`
> In the following information, `{single layer description}`:
> `{labeled JSON block(s)}`

The layer descriptions state that `views` contains three-view information, `shape` contains solid information, `metrics` contains measurement information, and `relations` contains spatial relation information. In the cumulative series, these descriptions follow the selected order. Text-only user prompts remove all references to the image; the question and selected description payloads are retained.

B.3 Serialization and empty fields

Each selected field is serialized as a labeled JSON block. The actual delimiters are the Unicode lenticular brackets U+3010 and U+3011; ordinary square brackets are used below for typographic convenience:

```
[views] { ... }
[shape] { ... }
[metrics] { ... given measurements only ... }
[relations] { ... }
```

For multiple solids, labels identify the solid and field, for example `[Solid 1's views]` and `[Solid 2's views]`. The `metrics.derived` branch is filtered before injection.

Block inclusion tests whether the selected value is `None`, rather than whether it has non-empty content. Thus an empty dictionary is rendered, whereas an absent or `None` value is skipped. A rendered empty Relations block can still change the header and surrounding instructions relative to the baseline. In the cumulative template, the introductory description lists the selected layers, rather than only those with non-empty content.

Images are downloaded from each problem's image URLs and encoded as base64. The same payload is reused across the image-present conditions for that problem. These conventions hold the source image fixed while varying its availability and the supplied text; they do not equate token length or wording across different interventions.

B.4 What the contrasts identify

Equations 1–3 answer distinct questions. A standalone gain asks whether supplying a component improves accuracy relative to the corresponding baseline. A conditional gain asks whether a component remains useful after the preceding components in the fixed order. The image contrast asks whether retaining the image changes accuracy under a specified information set and prompt series.

Overlap and order. The conditional contribution of Givens after Views and Shape need not equal its standalone contribution. Measurements may already be present in both preceding components. Similarly, a small Relations increment is conditional on the preceding descriptions and on its non-empty coverage. These contrasts are not order-invariant allocations of the total gain, and the evaluated subsets do not identify every interaction.

Net gains and answer transitions. For any comparison on the same N problems, let n_{ab} denote the number whose correctness is $a \in \{0, 1\}$ before and $b \in \{0, 1\}$ after the intervention. Then

$$A_{\text{after}} - A_{\text{before}} = \frac{n_{01} - n_{10}}{N}. \tag{4}$$

A positive gain is a net improvement: it subtracts newly introduced errors from corrected errors. Even the normalized quantity

$$\frac{A_{\text{after}} - A_{\text{before}}}{1 - A_{\text{before}}} = \frac{n_{01} - n_{10}}{n_{00} + n_{01}} \tag{5}$$

is not the fraction of baseline errors repaired, which would be $n_{01}/(n_{00} + n_{01})$. Residual error after the intervention is $(n_{00} + n_{10})/N$ and can include newly introduced failures. These identities explain the accuracy contrasts; they do not report an additional transition analysis.

Image utility and residual failure. A near-zero image contrast is an aggregate property. Gains on some problems can offset losses on others, so it does not establish that images are dispensable for every problem. Likewise, failure with full GeoText does not locate a failure inside a particular internal stage. Incomplete or inaccurate descriptions, interpretation of the supplied representation, geometric inference, calculation, and answer production remain possible contributors.

Scope of attribution. The interventions vary descriptive content together with its textual presentation. The evaluated design does not separate the benefit of geometric facts from every effect of prompt length, formatting, or wording. It supports comparisons between the specified input configurations, rather than a pure semantic treatment effect or a mechanistic decomposition of the model. No equivalence test or repeated-run noise estimate is used to interpret a small observed contrast as exact equality.

C Benchmark Composition and Description Availability

C.1 Composition and coverage

OrthoMind-Bench contains 5,357 Chinese K–12 solid-geometry problems. Table 5 reports its principal composition. Three-view drawings are the largest group, but the collection also contains solid diagrams and nets. Open-form questions are the majority, so scoring must accommodate answers beyond option letters.

Table 5: **Benchmark composition.** Percentages are rounded to one decimal place. The two batches have disjoint identifiers; they are not difficulty groups.

Attribute	Category	Problems	Share
Figure type	Three-view drawing	4,296	80.2%
	Solid diagram	638	11.9%
	Net	374	7.0%
	Other	49	0.9%
Answer format	Open form	4,024	75.1%
	Multiple choice	1,333	24.9%
Geometric entities	One	5,110	95.4%
	Two	99	1.8%
	Three or more	148	2.8%
Batch	1,420-problem batch	1,420	26.5%
	3,937-problem batch	3,937	73.5%

Table 6 shows that component availability is heterogeneous. The larger batch accounts for 73.5% of the benchmark but has only 9.4% non-empty Relations coverage. Aggregate Relations gains therefore combine problems with explicit relational content and problems with little or no such

Table 6: **Non-empty GeoText coverage by batch.** Coverage concerns stored content, not correctness, completeness, or the fraction of tasks requiring a component.

Batch	Views	Shape	Givens	Relations
1,420 problems	99.9%	100.0%	97.9%	100.0%
3,937 problems	81.5%	99.4%	87.8%	9.4%
All 5,357 problems	86.4%	99.5%	90.5%	33.4%

content in that field. Moreover, relations can already be expressed in Shape or Givens. Neither a small gain nor an empty Relations field establishes that a problem lacks relational demands.

Non-empty coverage also cannot validate annotation accuracy. GeoText is produced with Gemini-3.1-Pro-Preview, and its labels describe the information supplied by that pipeline. Full support selects all available fields; it is not a guarantee that every relevant fact has been captured.

C.2 Automatic composition labeling

A separate DeepSeek labeling pass characterizes problem composition using the API model alias `api_naci_default_deepseek-v4-pro`. Its input consists of the question, the `shape.cn` glosses joined across solids, and the `metrics.given.cn` glosses joined across solids. Empty description lines are omitted. The labeling input does not include `metrics.derived`. Nevertheless, because it uses retained GeoText descriptions, it is not an independent audit of their correctness or of target-related information.

The pass requests a JSON record with the fields in Table 7. The automatic difficulty labels are 46.7% basic, 49.6% medium, and 3.7% hard, with a mean estimated solution length of 2.70 steps. These are annotator judgments, not human-validated difficulty measurements or observed model reasoning lengths.

Table 7: **Composition-label schema.** Lists can contain multiple labels. Difficulty and solution length are estimated by the annotator.

Field	Requested output
`geometry_type`	A solid or construction type from the vocabulary below.
`dimension`	A list drawn from `views`, `shape`, `metrics`, and `relations`.
`category`	A list of specific test points from the vocabulary below.
`difficulty`	`basic`, `medium`, or `hard`.
`has_projection`	Boolean indicating point, line, or plane projection.
`has_sphere`	Boolean indicating a sphere or inscription/circumscription involving one.
`num_steps`	Integer estimate of the number of steps in a standard solution.
`summary`	One sentence describing the core task.

System instructions, English rendering. The annotator is instructed to act as an expert in classifying solid-geometry problems, classify the supplied problem, and output only a JSON object with the fields in Table 7, without Markdown fences or surrounding commentary. Its labeling criteria are as follows.

Solid vocabulary. Allowed labels are `Prism`, `Pyramid`, `Frustum`, `Cylinder`, `Cone`, `Sphere`, `ConeFrustum`, `YangMa`, `BiNao`, `ChuMeng`, `Union`, `Subtract`, `Cut`, `Truncate`, `HalfShape`, `QuarterShape`, `OctantShape`, `CubeStack`, `Revolve`, `CubeUnfold`, and `Other`. Prisms include cubes and cuboids; pyramids include tetrahedra. The construction labels distinguish joining, removal, sectioning, truncation, fractional solids, cube stacks, solids of revolution, and nets. The three traditional solid names are retained as labels.

Dimension criteria. `views` covers identifying projected shapes, lines, and labels, selecting or matching views, reading dimensions, and counting visible elements. `shape` covers recovering solid structure, interpreting nets, and analyzing stacks. `metrics` covers extracting measurements,

binding them to elements, and tasks involving volume, area, distance, angle, ratios, or single values. `relations` covers spatial constraints, projections, inscription/circumscription, and proofs. The instructions emphasize the task's central demand: a numerical answer alone does not require assigning `metrics` as the primary dimension when the main difficulty is relational.

Test-point vocabulary. Allowed categories are `view_select`, `view_match`, `view_readout`, `view_count`, `shape_identify`, `shape_reconstruct`, `unfold`, `stack_analysis`, `volume`, `surface_area`, `lateral_area`, `distance`, `angle`, `ratio`, `single_value`, `position_relation`, `projection`, `inscribed`, and `proof`. Here `single_value` denotes a single length or angle, and `position_relation` includes parallel, perpendicular, skew, and coplanar relationships.

Difficulty criteria. Basic problems provide views and dimensions directly and require one formula or approximately one to two steps. Medium problems require reconstructing the solid or computing an intermediate quantity, usually in two to three steps. Hard problems involve composite or sectioned solids, spheres, or multistep spatial reasoning, typically four or more steps. These criteria are heuristic and can overlap.

User template, English translation. The corresponding English user template is:

```
Problem: {question}
Solid description: {all shape.cn glosses, joined}
Given measurements: {all metrics.given.cn glosses, joined}
```

Execution and parsing. The labeling configuration uses a 4,096-token output limit, four worker threads, and up to three retries. Post-processing removes JSON fences when present and attempts JSON parsing. If direct parsing fails, it attempts to parse the span from the first opening brace to the last closing brace. Failed annotation attempts are recorded as error objects.

D Model Execution and Answer Scoring

D.1 Models and execution settings

The evaluated VLMs are Qwen3.5-2B, Qwen3.5-4B, Qwen3.5-9B, Qwen3.5-27B, GLM-4.5V, and InternVL3.5-8B. Table 8 reports the evaluation configuration. Each problem–condition pair has one evaluation run; retries handle request failures and do not constitute repeated-seed evaluation.

Table 8: **VLM evaluation configuration.** These settings describe the submitted requests.

Setting	Value
Temperature	0
Top-p	0.9
Maximum output tokens	10,000
Streaming	Disabled
Thinking flag	`enable_thinking = false`
Runs per problem–condition pair	1
Worker threads	40
Pause between problems	2 seconds
Request retries / backoff	3 / 10 seconds

Temperature zero does not by itself establish bitwise determinism across hardware or serving configurations. Without repeated runs, the experiment does not quantify a run-to-run noise floor. Small differences are therefore described as observed accuracy contrasts, without treating them as proof of statistical equivalence.

D.2 Extraction and rule-based matching

The evaluator first searches for the final-answer span using the following regular expression with DOTALL and case-insensitive matching:

```
<answer>\s*(.+?)\s*</answer>
```

If a match is found, its captured content is used; otherwise, the raw response is retained. The candidate and reference answer are compared directly and after normalization. Normalization removes whitespace, including full-width spaces; maps full-width parentheses and commas to their ASCII counterparts; and replaces the LaTeX command \pi with the corresponding mathematical symbol π.

The rule-based result is correct when either the direct strings or their normalized forms agree. Multiple-choice responses use the same matching function, with the option letter as the reference. There is no separate semantic option-selection branch. This procedure is a string matcher, not a symbolic equivalence engine: mathematically equivalent expressions can remain different after normalization.

D.3 Answer-level rechecking and interpretation

Qwen3-30B-A3B is used for answer-level rechecking in the evaluation pipeline. This step concerns agreement between a candidate answer and the reference answer; it does not validate intermediate reasoning or certify the geometric descriptions.

Final-answer accuracy supports the input comparisons in Section 2.3. It does not establish the validity of a solution trace, identify a particular reasoning mechanism, or determine whether a failure originates in a supplied description. Accordingly, errors under full GeoText are reported as residual errors under the specified support, rather than as a measured category of computational failure.

E Numerical Results for the Diagnostic Interventions

This appendix records the numerical results used in Section 3. All aggregate tables concern $N = 5{,}357$ problems; the final table separates the two batches for the available Qwen3.5-27B comparison. Accuracies and error rates are percentages, while accuracy differences are percentage points (pp). Component definitions and prompt-series notation follow Section 2.3.

E.1 Cumulative accuracies with images

Table 9 reports the image-present cumulative sequence. The selected subsets are $U_0 = \varnothing$, $U_1 = \{\mathrm{V}\}$, $U_2 = \{\mathrm{V}, \mathrm{S}\}$, $U_3 = \{\mathrm{V}, \mathrm{S}, \mathrm{G}\}$, and $U_4 = \mathcal{C}$. Each entry is $100A_m^{(\mathrm{cum})}(U_k, 1)$.

Table 9: **Image-present cumulative accuracy (%).** Columns progressively supply Views, Shape, Givens, and Relations. All values are rounded to two decimals.

Model	U_0	U_1	U_2	U_3	U_4
Qwen3.5-2B	31.44	35.42	51.56	50.92	52.16
Qwen3.5-4B	43.35	48.63	58.13	59.85	60.00
Qwen3.5-9B	47.82	55.85	63.30	65.30	64.25
Qwen3.5-27B	50.63	58.47	65.76	66.70	67.37
GLM-4.5V	34.88	40.35	50.16	51.88	53.82
InternVL3.5-8B	24.11	35.72	55.24	56.11	57.16

The successive differences reproduce the conditional gains in Table 10. Two increments are negative: -0.64 pp for adding Givens to Qwen3.5-2B and -1.05 pp for adding Relations to Qwen3.5-9B. These are the non-monotonic cases discussed in Section 3.3; they are observed aggregate changes, without an inferred attribution to a particular failure mechanism.

E.2 STANDALONE AND CONDITIONAL COMPONENT GAINS

Table 10 supplies all entries in Figure 2. Standalone gains use the single-component series and its baseline; conditional gains use successive conditions within the cumulative series. As specified in Appendix B, identical information subsets need not have identical introductory wording across series.

Table 10: **Component gains with images present (pp).** Qwen columns identify parameter counts within the Qwen3.5 family; GLM denotes GLM-4.5V and InternVL denotes InternVL3.5-8B. Conditional order is Views → Shape → Givens → Relations.

Comparison	Component	2B	4B	9B	27B	GLM	InternVL
Standalone	Views	3.98	5.28	8.03	7.84	5.47	11.61
	Shape	16.72	15.94	14.42	13.34	13.15	28.74
	Givens	10.71	10.45	9.64	8.56	8.67	19.37
	Relations	0.82	1.04	1.63	2.25	−0.50	4.60
Conditional	Views	3.98	5.28	8.03	7.84	5.47	11.61
	Shape	16.14	9.50	7.45	7.29	9.81	19.52
	Givens	−0.64	1.72	2.00	0.94	1.72	0.87
	Relations	1.24	0.15	−1.05	0.67	1.94	1.05

The Shape row is largest among standalone components for every model, and its conditional row remains positive throughout. The Givens row decreases from standalone to conditional evaluation in every model. These statements describe the reported comparisons, rather than additive allocations of total performance.

Precision of the reconstructed Shape accuracies. The $I{+}S$ column of Table 2 adds the reported image-present baseline to the reported standalone Shape gain. Both inputs are rounded to two decimals; their sum can consequently differ from a directly rounded condition accuracy by 0.01 pp. The column is marked as reconstructed in the main-table caption. Neither integer counts of correct answers nor more precise accuracies are inferred from these rounded summaries.

E.3 IMAGE CONTRASTS, TOTAL GAINS, AND REMAINING ERROR

Table 11 supplies the endpoints of Figure 3 and the coordinates of Figure 4. It also reports the net full-GeoText accuracy gain discussed in Section 3.3. With accuracy expressed as a fraction, the quantities are

$$d_{m,0} = 100\Big[A_m^{(\mathrm{cum})}(\varnothing, 1) - A_m^{(\mathrm{cum})}(\varnothing, 0)\Big], \tag{6}$$

$$d_{m,\mathrm{All}} = 100\Big[A_m^{(\mathrm{cum})}(\mathcal{C}, 1) - A_m^{(\mathrm{cum})}(\mathcal{C}, 0)\Big], \tag{7}$$

$$g_{m,I} = 100\Big[A_m^{(\mathrm{cum})}(\mathcal{C}, 1) - A_m^{(\mathrm{cum})}(\varnothing, 1)\Big], \tag{8}$$

$$g_{m,T} = 100\Big[A_m^{(\mathrm{cum})}(\mathcal{C}, 0) - A_m^{(\mathrm{cum})}(\varnothing, 0)\Big], \tag{9}$$

$$e_{m,I} = 100\Big[1 - A_m^{(\mathrm{cum})}(\mathcal{C}, 1)\Big], \tag{10}$$

$$e_{m,T} = 100\Big[1 - A_m^{(\mathrm{cum})}(\mathcal{C}, 0)\Big]. \tag{11}$$

Figure 3 compares $d_{m,0}$ with $d_{m,\mathrm{All}}$. Figure 4 plots $(d_{m,\mathrm{All}}, e_{m,I})$, so it displays two measurements from the same full-support setting rather than successive stages of an error decomposition. The tabulated endpoint comparison does not specify the image contrasts at intermediate subsets.

Net gains and error transitions. The gain $g_{m,I}$ ranges from 16.43 to 33.05 pp and equals the sum of the four conditional increments, because successive cumulative differences telescope. This arithmetic identity does not make the increments contributions of disjoint error types. As formalized

Table 11: **Image utility, full-support gain, and residual error.** $d_{m,0}$ and $d_{m,\mathrm{All}}$ are image contrasts before and after full GeoText; $g_{m,I}$ and $g_{m,T}$ are its net gains with and without images. These four columns use pp; $e_{m,I}$ and $e_{m,T}$ are error rates (%). All entries are computed from the reported endpoint accuracies.

Model	$d_{m,0}$	$d_{m,\mathrm{All}}$	$g_{m,I}$	$g_{m,T}$	$e_{m,I}$	$e_{m,T}$
Qwen3.5-2B	19.22	8.16	20.72	31.78	47.84	56.00
Qwen3.5-4B	22.79	−5.41	16.65	44.85	40.00	34.59
Qwen3.5-9B	25.36	−5.12	16.43	46.91	35.75	30.63
Qwen3.5-27B	27.89	−1.87	16.74	46.50	32.63	30.76
GLM-4.5V	14.97	−0.20	18.94	34.11	46.18	45.98
InternVL3.5-8B	10.49	6.93	33.05	36.61	42.84	49.77

in Appendix B.4, a net gain subtracts newly introduced errors from corrected errors. No per-problem transition counts or error-type frequencies are inferred from the aggregate tables.

Relation between GeoText gains and image contrasts. The text-only gains satisfy $g_{m,T} > g_{m,I}$ for every model. Algebraically,

$$d_{m,\mathrm{All}} - d_{m,0} = g_{m,I} - g_{m,T}. \tag{12}$$

The reduction in image benefit and the larger text-only gain are therefore two views of the same endpoint comparison, not independent pieces of evidence.

Remaining errors with and without images. The image-present residual range is 32.63–47.84%, and the image-absent range is 30.63–56.00%. The latter confirms that removing the image does not eliminate the remaining performance gap. These are error rates under the specified descriptions and evaluation settings, rather than estimates of a reasoning-only failure component.

E.4 Relations in the two batches

Table 12 reports the available batch-level comparison for Qwen3.5-27B. The contrast adds Relations after Views, Shape, and Givens, with images retained. Non-empty coverage measures whether the Relations field contains content, not whether a problem requires relational reasoning.

Table 12: **Qwen3.5-27B: conditional Relations gain by batch.** The two batches have disjoint problem identifiers and differ in composition as well as Relations coverage.

Evaluation set	Problems	Relations coverage (%)	Gain (pp)
1,420-problem batch	1,420	100.0	0.74
3,937-problem batch	3,937	9.4	0.64
Combined	5,357	33.4	0.67

The size-weighted mean of the rounded batch gains is approximately 0.67 pp, consistent with the reported combined result. The small gain in the fully Relations-covered batch shows that empty content is not a sufficient explanation for this model's small conditional increment. Because the batches differ and earlier components can express relational information, this comparison does not isolate a causal effect of coverage. It also provides no batch-level claim for the other five models.