NoReGeo: Non-Reasoning Geometry Benchmark

Irina Abdullaeva[1,2], Anton Vasiliuk[1], Elizaveta Goncharova[1,3], Temurbek Rahmatullaev[1,4], Zagorulko Ivan[5], Maxim Kurkin[1, 6], Andrey Kuznetsov[1,2]

[1]FusionBrain Lab, Russia
[2]Research Center of the Artificial Intelligence Institute, Innopolis University, Innopolis, Russia
[3]HSE University, Russia
[4]Lomonosov Moscow State University, Russia
[5]Central University, Russia
[6]Applied AI Institute, Moscow, Russia
abdullaeva@fusionbrainlab.com, goncharova@fusionbrainlab.com, kuznetsov@fusionbrainlab.com

Abstract

We present NoReGeo, a novel benchmark designed to evaluate the intrinsic geometric understanding of large language models (LLMs) without relying on reasoning or algebraic computation. Unlike existing benchmarks that primarily assess models' proficiency in reasoning-based geometry-where solutions are derived using algebraic methods-NoReGeo focuses on evaluating whether LLMs can inherently encode spatial relationships and recognize geometric properties directly. Our benchmark comprises 2,500 trivial geometric problems spanning 25 categories, each carefully crafted to be solvable purely through native geometric understanding, assuming known object locations. We assess a range of state-of-the-art models on NoReGeo, including frontier models like GPT-4, observing that even the most advanced systems achieve an overall maximum of 65% accuracy in binary classification tasks. Further, our ablation experiments demonstrate that such geometric understanding does not emerge through fine-tuning alone, indicating that effective training for geometric comprehension requires a specialized approach from the outset. Our findings highlight a significant gap in current LLMs' ability to natively grasp geometric concepts, providing a foundation for future research toward models with true geometric cognition.

Code — https://github.com/FusionBrainLab/NoReGeo

Introduction

Although modern LLMs excel at symbolic reasoning, they still treat even the simplest spatial relations as formal reasoning problems, producing multi-step chains of thought instead of relying on an intuitive sense of geometry. Ask whether two line segments intersect, and many models unfold a miniature proof rather than making a direct geometric judgment. This gap between symbolic reasoning and intuitive spatial understanding becomes especially limiting in time-critical settings — CAD engines updating thousands of vertices per frame, robots refining grasp trajectories in milliseconds, or geospatial systems performing rapid visibility checks. In such environments, approximate geometric intuition is far more valuable than multi-line chains of thought. Yet current models lack native geometric understanding and must simulate reasoning even for the simplest relations.

Existing geometry benchmarks primarily emphasize complex proofs or multi-step algebraic reasoning (Lu et al. 2021; Kazemi et al. 2024), conflating two skills: (i) identifying relevant spatial facts and (ii) executing symbolic derivations. To isolate the first skill, we introduce **NoReGeo**, a benchmark of 2,500 trivially solvable geometry problems across 25 categories. Each item can be answered directly from point locations, without auxiliary constructions, theorems, or lengthy CoT. Every problem appears in both text-only form and a paired diagram, enabling controlled comparisons between text-based LLMs and multimodal vision-language models (VLMs).

We evaluate more than 45 frontier and open-source models under both modalities. Even the strongest open-source VLM achieves only $\sim 55\%$ accuracy (Phi3.5-Vision), and the best proprietary model reaches $\sim 65\%$ — well below human performance of $\sim 74.5\%$ on the same multiple-choice tasks. These results expose a substantial gap in basic spatial intuition. Further analysis shows that fine-tuning alone does not confer geometric competence, whereas a simple linear probe on a frozen vision encoder solves the tasks almost perfectly — suggesting that geometric features are present in embeddings but are not accessed by current LLM architectures or training regimes.

Our contributions are as follows:

1. We motivate and formalize the concept of *native geometric understanding* as a core capability required for spatially intensive applications.
2. We introduce **NoReGeo**, the first benchmark explicitly designed to test this ability without chain-of-thought or algebraic computation, comprising 2 500 items across 25 categories in both text and image formats.
3. We provide a comprehensive evaluation of over 45 state-of-the-art LLMs and VLMs, showing that all fall short of human-level performance on elementary geometric tasks.
4. Through fine-tuning and linear-probing studies, we demonstrate that geometric knowledge exists latently in vision encoders but fails to naturally emerge in current LLM training paradigms.

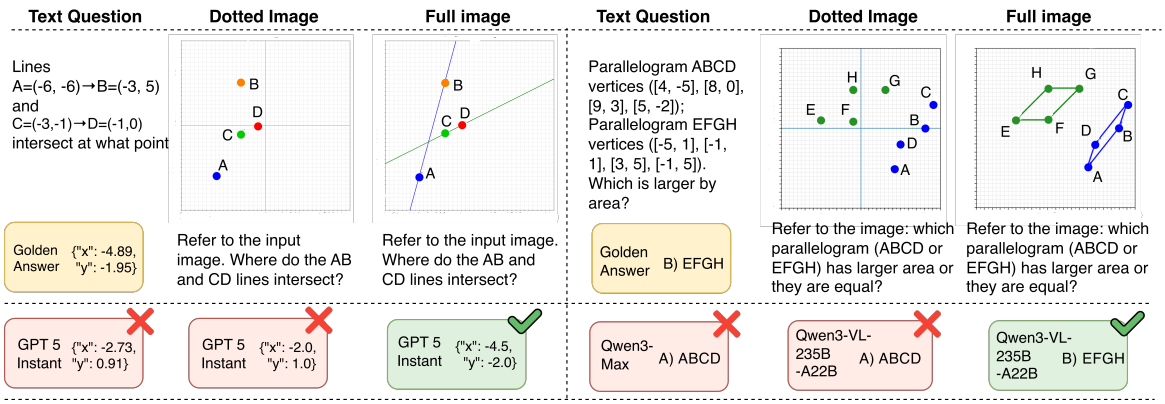

Figure 1: Evaluation samples from NoReGeo benchmark. Each problem is shown in three formats – (a) text-only, (b) text with dotted-image (points only), and (c) text with full-image (points plus connecting lines) – together with the golden answer (yellow) and the model's prediction.

Overall, our findings point to an urgent open problem: ***bridging the gap between symbolic reasoning and true geometric cognition in future foundation models.***

Related Work

Spatial reasoning in vision–language models. Vision–language systems must ground textual concepts such as *left of* or *bigger than* in a spatial frame of reference. Synthetic VQA benchmarks like CLEVR isolate this ability by rendering scenes of coloured shapes and asking queries that combine Boolean logic with coarse spatial predicates (Abraham, Alirezaie, and Raedt 2024). Subsequent datasets move toward natural imagery: SPATIALSENSE adversarially mines object pairs (e.g. *person–bench*) to evaluate relative positioning (Yang, Russakovsky, and Deng 2019), while 3DSRBENCH embeds similar relations in RGB-D scans of indoor environments (Ma et al. 2025). Interactive navigation tasks extend spatial grounding to embodied settings: BabyAI and ALFRED require locating, manipulating, and placing objects in simulated rooms (Chevalier-Boisvert et al. 2019; Shridhar et al. 2020), and Room-to-Room (R2R) tests natural-language navigation in realistic 3-D reconstructions (Hong et al. 2025). Together, these datasets show that VLMs handle *qualitative* spatial relations reasonably well, yet they provide little evidence that models encode the **fine-grained geometric attributes** — distances, angles, midpoints, coordinate relations — needed for the intuitive geometric understanding.

Geometry word-problem benchmarks. A parallel line of work examines whether models can solve textbook geometry questions that pair a diagram with natural language. Benchmarks such as GEOMETRY3K (Lu et al. 2021), PGPS-9K (Zhang, Yin, and Liu 2023), MathVista (Lu et al. 2024), MathVerse (Zhang et al. 2024b), and GEOMVERSE (Kazemi et al. 2024) collect thousands of K–12 geometry problems, while GEOEVAL adds a calibrated difficulty ladder and reports that even math-tuned LLMs plateau at $\approx 55\%$ accuracy on "regular" items and $< 10\%$ on Olympiad-level ones (Zhang et al. 2024a). Solving these tasks typically requires *multi-step reasoning*: interpreting diagrams, identifying theorems, and chaining symbolic deductions. As a result, existing evaluations emphasize algebraic manipulation and proof-style solution pipelines — often supported by chain-of-thought prompting, external tools, or full symbolic proof synthesis (e.g. AlphaGeometry (Trinh et al. 2024)) — making it difficult to determine whether models possess any *native geometric perception* independent of reasoning.

Skill-specific probes. To obtain a more targeted diagnostic signal, several recent efforts isolate individual geometric skills. PLANQA presents floor-plan layouts as JSON and asks questions about visibility or shortest paths (Rodionov et al. 2025). GEOMREL focuses on detecting equal segments and angles before any numeric computation (Wang et al. 2025), and GEOGRAMBENCH converts Asymptote-style code into natural-language questions of varying abstraction (Luo et al. 2025b). These benchmarks remove some confounds but still require models to parse code-like input formats or engage in subtle inference — leaving open the question of whether models can make *direct, perception-level* geometric judgments.

The proposed cross-modal NOREGEO benchmark targets **single-step, school-level geometry questions** (midpoints, area comparisons, collinearity, symmetry tests, and similar micro-skills) that can be answered instantaneously by anyone with basic geometric intuition with no auxiliary constructions, theorem recall, or multi-hop reasoning. By providing both text-only and vision-augmented variants under a shared output format, NOREGEO enables controlled comparisons between LLMs and VLMs and reveals whether either modality supplies genuine geometric understanding. Because the tasks are *synthetic yet curriculum-aligned*, they are unlikely to appear verbatim in pre-training corpora, ensuring that NoReGeo is a genuine test of latent geometric understanding rather than memorization.

Thus, our benchmark complements prior work by stripping away reasoning scaffolds and focusing on the bedrock

geometric knowledge that more complex systems implicitly assume.

Benchmark Motivation and Scope

The NoReGeo is designed as a cross-modal geometry-based benchmark oriented to probe the **foundational geometric competence** of modern LLMs and VLMs. The benchmark consists of short prompt–answer pairs in elementary geometry. Each problem is posed as a one-shot query: the model receives a single prompt and must return an answer immediately, without any chain-of-thought or intermediate steps. The benchmark provides two prompt modalities — text-only questions and their corresponding image-based versions (where the images are presented in the so-called dotted and full versions). The example of samples from the NoReGeo can be found in Figure 1.

Each problem has a ground-truth answer that is either a numeric value (for quantitative questions, typically an integer or simple fraction; e.g. 90) or a categorical label (for qualitative classification questions; e.g. "acute" to describe an angle type). There are no elaborate proofs or explanations required – the output is a single final answer. The evaluation metric is straightforward accuracy for multiple-choice questions and soft accuracy (within the [-0.5, 0.5] interval) for numeric ones.

High-Level Taxonomy of Tasks

During benchmark creation, we followed the typical secondary school geometry curriculum, our benchmark effectively covers content taught from roughly higher school (National Governors Association Center for Best Practices and Council of Chief State School Officers 2010; National Council of Teachers of Mathematics 2000). We have included some foundational topics that are introduced in middle school (for example, basic angle facts or simple constructions from early secondary years) as well as the full suite of high-school geometry topics (Euclidean proofs, circle theorems, introductory trigonometry, etc.) The detailed taxonomy of NoReGeo is given in Table 1.

The benchmark tasks are categorized into three types: *Classification*, *Numeric*, and *Unstable*. Classification tasks (C-) involve multiple-choice questions, such as identifying polygon areas or symmetry. Numeric tasks (N-) require numeric values like coordinates or lengths, while Unstable tasks (U-) involve binary decisions that can change with minimal input variation.

Dataset Construction

Building on the taxonomy described above, we developed a concise pipeline for synthetic data generation. Each benchmark item is provided either in text-only form or as a multimodal (vision–text) variant combining text with an image.

All items in the dataset follow these design rules:

- Every vertex uses integer coordinates in the range $[-20, 20]$. When an image is present, the points lie on a Cartesian grid.
- In text-only prompts, points are denoted by uppercase letters with their coordinates, e.g. $A = (2, 1)$.

The multimodal format has two variants: (i) an image showing only labeled points (without coordinates labels), with edges implied by the text; or (ii) an image of the complete figure, where the text provides only the question (and any answer options) without listing coordinates. The sample questions are give below.

Text-only. Lines $A = (2, 1) \to B = (3, 0)$ and $C = (-8, -1) \to D = (9, 0)$ intersect at what point?

Multimodal. Refer to the input image. Where do lines AB and CD intersect?

We split the questions into *vision-only* and *text-only* tasks to minimize the influence of the textual prompt on the vision-based tasks, ensuring that coordinates must be read solely from the image that is a significant challenge in modern multimodal benchmarks (Chen et al. 2024).

As shown in Table 1, the benchmark maintains a slight emphasis on basic coordinate geometry over symmetry tasks. Notably, there is a significant majority of classification tasks compared to numerical tasks. This design choice prioritizes the evaluation of a model's ability to recognize geometric properties and apply definitions over pure computation. All problems are solvable through the direct application of fundamental formulas and definitions, making the benchmark a robust tool for assessing core geometric understanding in AI models across both text and vision modalities.

Experiments

In this section, we evaluate whether LLMs and VLMs can natively perceive geometric structures and relationships across varying text-to-visual information ratios using our NoReGeo benchmark.

Experimental Setup

Models and Implementation. We selected a broad range of state-of-the-art LLMs and VLMs, including both proprietary and open-source models, to capture representative examples of current capabilities. Our model selection also allows us to analyze trends across model generations and explore the effect of model scale on geometric and multimodal reasoning. In total, we evaluated over 45 models. Proprietary API-based models include GPT-4.1, GPT-4.1-Mini, and GPT-4.1-Nano. Open-source models span the Qwen series (Qwen2, Qwen2.5 (Team 2024) and Qwen3 (Yang et al. 2025), along with multimodal variants Qwen2-VL (Wang et al. 2024) and Qwen2.5-VL (Bai et al. 2025)), the LLaMA-3.1 series (Grattafiori et al. 2024), and Mistral models (Jiang et al. 2023) (including versions with math-specific pretraining), among others. Additionally, we evaluated math-specific VLMs, including the G-LLaVA (Gao et al. 2023), URSA (Luo et al. 2025a), Math-LLaVA (Shi et al. 2024), and Multimath-7B-LLaVA-v1.5 (Peng et al. 2024), to contrast general-purpose and specialized models.

To ensure consistency across models, we standardized generation settings: fixed random seed, temperature set to 0.6, and a maximum output length of 2048 tokens. For instruction-tuned models, we used their native chat templates and applied a unified system prompt.

Type	Category	Task	ID	Type	Sample Question
Class.	Area comparison	parallelogram_size	C-ACM-PST	MC	Parallelograms ABCD and EFGH (vertices given). Which has larger area?
		triangle_size	C-ACM-TST	MC	Triangles ABC and DEF. Which has larger area?
	Basic coordinate tasks	collinearity	C-BCT-CT	MC	Are points A, B, and C collinear?
	Symmetry	shape_symmetry	C-SYM-SST	MC	Polygon (vertices given): symmetric about Y-axis?
Numeric	Basic coordinate tasks	midpoint	N-BCT-MT	Coord	Find midpoint of segment AB.
		parallelogram_area	N-BCT-PAT	Num	Parallelogram ABCD: find area.
		parallelogram_perim	N-BCT-PPT	Num	Parallelogram ABCD: find perimeter.
		triangle_area	N-BCT-TAT	Num	Triangle ABC: find area.
		triangle_perim	N-BCT-TPT	Num	Triangle ABC: find perimeter.
	Elementary calc.	intersection	N-ECL-IT	Coord	Lines AB and CD: find intersection point.
		segment_length	N-ECL-SLT	Num	Segment AB: find length.
	Geometric transform.	reflection	N-GTR-RT	Coord	Reflect point P across X-axis. Find coordinates.
		rotation_point	N-GTR-RPT	Coord	Rotate point P 90° about origin.
	Remarkable triangle lines	bisector	N-RLT-BT	Coord	Triangle ABC: find B-angle bisector intersection.
	Simple circle properties	inner_circle_center	N-SCP-ICCT	Coord	Triangle ABC: find incircle center.
		inner_circle_radius	N-SCP-ICRT	Num	Triangle ABC: find incircle radius.
		outer_circle_center	N-SCP-OCCT	Coord	Triangle ABC: find circumcenter.
		triangle_type	N-SCP-TTT	Num	Triangle ABC: find triangle type.
Unstable	Circle properties	circle_diameter	U-CPR-CDT	MC	Circle with center O, radius = 3; is AB a diameter?
	Parallelism and Perpendicularity	parallel_lines	U-PAP-PLT	MC	Are lines AB and CD parallel?
		perpendicular	U-PAP-PT	MC	Are segments AB and CD perpendicular?
		right_angle	U-PAP-RAT	MC	Is angle ABC a right angle?
	Remarkable triangle lines	special_lines	U-RLT-SLT	MC	Triangle ABC with segment from B to D. Identify the segment.
	Simple circle properties	semicircle_triangle	U-SCP-STT	MC	Is triangle ABC inscribed in a semicircle?
	Symmetry	symmetry	U-SYM-ST	MC	Are points P and Q symmetric about line y=x?

Table 1: Overview of the NoReGeo benchmark tasks, organized by question type, geometric category, and specific task. Each task is identified by a structured ID code consisting of: (i) a single-letter type prefix (C: Classification, N: Numeric, U: Unstable), (ii) a three-letter category code, and (iii) a short task name. The benchmark covers a broad range of geometric reasoning tasks including area and perimeter comparison, coordinate calculations, symmetry, triangle centers, and more.

Evaluation Scheme. We focus on each model's direct-answer capability - its ability to solve tasks without producing intermediate reasoning steps. To enforce structured output and prevent unsolicited reasoning, we applied a structured generation approach. Each task specifies a JSON-formatted response template based on the expected answer type (e.g., multiple choice, numeric value, or coordinate point). This structure is communicated to the model through a structure prompt appended to each question.

We implemented this setup using the Outlines (Willard and Louf 2023) and xgrammar (Dong et al. 2024) libraries, which convert expected JSON structures into regular expressions. These are compiled into finite state machines that bias model generation by modifying logits. For efficient model serving, we used the VLLM library (Kwon et al. 2023).

Evaluation Metrics & Policy. We evaluated the capabilities of LLMs and VLMs by comparing generated and reference answers using the accuracy metric. For multiple-choice tasks, answers were considered to match if they were an exact match for one of the generated answer options. For numerical and coordinate answers, we defined a tolerable error interval of 0.5, identical to the grid step in visualizations of geometric problems. Numerical answers and point coordinates were considered correct if they met the following criteria: a) they were valid numbers; b) they fell within the interval [reference answer - 0.5, reference answer + 0.5]. Regarding point coordinates, both coordinates of the answer point also had to be correct according to the numerical answer criteria mentioned above. To gain a more detailed understanding of error magnitude in cases involving numerical answers, we calculated regression metrics (mean squared error, MSE) for tasks where the answer was either a number or a point.

If the answer did not match the required format — if it

	Classification			Numeric					Unstable				
	ACM	BCT	SYM	BCT	ECL	GTR	RLT	SCP	CPR	PAP	RLT	SCP	SYM
Text													
Qwen2.5-3B-In.	**69.2**	44.0	54.0	3.9	**32.3**	0.0	*98.5*	*32.2*	*53.0*	47.8	34.5	0.0	46.5
Qwen2.5-7B-In.	24.5	52.0	57.0	5.6	12.7	**92.0**	**99.0**	21.2	6.0	21.3	22.0	7.0	47.0
Qwen3-4B	56.5	61.0	52.0	5.0	*17.0*	0.0	40.0	**36.2**	52.0	52.7	31.0	49.0	*89.0*
Qwen3-8B	58.0	*73.0*	67.0	**6.8**	10.7	0.0	4.0	25.8	52.0	*53.0*	**47.0**	*64.0*	**91.0**
Mistral-Small-In.	57.5	56.0	59.0	3.8	12.0	0.0	3.0	14.5	52.0	45.3	32.0	50.0	79.0
LLaMa-3.1 8B-In.	64.5	51.0	*96.0*	1.6	9.3	0.0	18.0	15.2	*53.0*	52.0	40.0	**67.0**	68.0
LLaMa-3.1 70B-In.	*67.0*	**78.0**	**97.0**	*6.4*	16.7	0.0	0.0	17.2	**97.0**	**53.7**	*43.0*	52.0	84.0
Text with dotted images													
Qwen2-VL-7B-In.	38.5	54.0	52.0	9.4	35.7	*50.5*	*50.0*	*57.2*	52.0	*50.0*	76.0	49.0	49.0
Qwen2.5-VL-7B-In.	40.5	51.0	47.0	3.4	12.3	0.0	1.0	31.5	52.0	48.3	92.0	43.0	50.0
InternVL2.5-8B	34.0	59.0	52.0	*18.6*	**60.7**	8.5	28.0	63.0	52.0	43.7	37.0	58.0	46.0
InternVL3-8B	*41.0*	53.0	52.0	2.8	5.3	0.0	17.0	39.2	52.0	55.7	**99.0**	51.0	54.0
LLaVA-Mini (LLaMA-8B)	20.5	*62.0*	59.0	20.8	*56.7*	**75.0**	**90.0**	43.5	53.0	47.0	29.0	52.0	39.0
MiniCPM-o-2.6	40.0	44.0	*54.0*	7.8	17.0	0.0	2.0	48.5	44.0	*50.0*	42.0	49.0	53.0
Phi-3.5-Vis.-In.	41.5	**73.0**	52.0	15.4	32.3	1.0	35.0	46.5	52.0	49.3	*93.0*	*55.0*	50.0
Human eval	**81.5**	70.0	**72.0**	**63.0**	0.0	5.0	50.0	**78.5**	**88.0**	**92.0**	89.0	**81.0**	**94.0**
Text with full images													
Qwen2-VL-7B-In.	*66.0*	88.0	90.0	**81.8**	**40.0**	*50.0*	*16.0*	36.0	**100.0**	55.7	71.0	*99.0*	**86.0**
Qwen2.5-VL-7B-In.	**66.5**	**100.0**	*99.0*	*79.0*	33.7	0.0	0.0	26.8	**100.0**	*63.0*	**98.0**	*99.0*	*62.0*
InternVL2.5-8B	53.5	96.0	67.0	75.6	32.7	1.5	10.0	**44.0**	73.0	50.7	41.0	88.0	60.0
LLaVA-Mini (LLaMA-8B)	24.5	52.0	57.0	5.6	12.7	**92.0**	**99.0**	21.2	6.0	21.3	22.0	7.0	47.0
MiniCPM-o-2.6	52.5	88.0	86.0	77.0	31.0	0.0	0.0	17.5	*97.0*	**64.0**	63.0	96.0	57.0
Phi-3.5-Vis.-In.	50.0	*99.0*	**100.0**	57.6	*39.3*	0.0	13.0	*41.8*	65.0	52.3	*94.0*	**100.0**	*62.0*

Table 2: Accuracy (%) of selected models on each benchmark task category across three setups: text-only, text with dotted and full images. **Bold** indicates best per setup, *italic* – second-best, and underlined shows overall best across all setups.

was not valid JSON, included an additional reasoning trail, or lacked the correct answer fields in JSON — we marked it as incorrect and awarded no credit, even if the answer was mathematically correct. This strict policy isolates a model's geometric competence from its propensity to reveal private reasoning.

Human Evaluation For the dotted image format, we also conducted a human evaluation to assess how accurately humans can solve these tasks. The human baseline was obtained using the Toloka platform. Annotators completed training, examination, and control tasks, with 10 tasks per page and 10 minutes allowed per page. Each task was answered by three crowd workers, and majority voting was used to aggregate responses. Participants were instructed not to use external resources; the only aid was the task's dotted-format overlay. The average annotator age was 39 years, with compensation of approximately $1 per page.

Main results

General performance. There is a significant disparity in the extent to which different models comprehend geometry and leverage cross-modal relations. Table 2 shows the average quality of a few representative LLMs and VLMs across task categories. Table 3 shows the average quality of problem solving across all evaluated models within task categories, taking into account the standard deviation. Based on these results, we draw several key conclusions.

Full visual context significantly boosts VLM performance. Models evaluated with text and full images consistently outperform both text-only and text with dotted images settings across nearly all task types and categories. For example, Qwen2.5-VL-7B-Instruct reaches 100% accuracy on several classification and unstable tasks (basic coordinate tasks, symmetry, circle properties) when provided with full image input, compared to much lower scores in the dotted-image setup. It also shows markedly improved performance on numerically intensive tasks such as basic coordinate tasks (79.0%) and elementary calculations (33.7%), which are typically challenging across the board. This strong overall trend is visualized in Figure 2, where most task categories exhibit substantial positive accuracy gaps favoring full images. The largest average gains occur in tasks involving curved shapes and global geometry, such as Area comparison (ACM), Basic Coordinates Tasks (BCT) and Circle Properties (CPR).

By contrast, linear or axis-aligned tasks, such as Parallelism, Perpendicularity, or Geometric Transformations (GTR), show minimal or no improvement between dotted and full visual input. This suggests that dots-only representations already encode sufficient information for solving simpler spatial alignment problems. A closer task-level breakdown further reinforces this distinction: while some tasks yield dramatic gains of +40–100% when full images are provided, others exhibit near-zero or even negative improvements.

Task category	Text-only	Text with dot images	Text with full images
Area comparison (ACM)	55.1 ± 19.1	33.6 ± 10.2	50.6 ± 23.5
Basic coordinate tasks (BCT)	32.4 ± 39.5	19.2 ± 25.9	53.4 ± 42.8
Circle properties (CPR)	59.4 ± 22.3	51.5 ± 4.6	71.0 ± 32.3
Elementary calculations (ECL)	26.1 ± 36.6	38.3 ± 35.9	33.7 ± 37.9
Geometric transformations (GTR)	12.1 ± 31.2	27.9 ± 36.6	18.6 ± 34.6
Parallelism and Perpendicularity (PAR)	51.6 ± 15.6	48.8 ± 9.1	49.7 ± 20.1
Remarkable lines of a triangle (RLT)	32.8 ± 30.9	57.3 ± 31.4	42.8 ± 37.0
Simple circle properties (SCP)	35.2 ± 34.4	47.6 ± 28.2	39.4 ± 30.9
Symmetry (SYM)	65.4 ± 19.6	50.0 ± 8.3	68.9 ± 27.5

Table 3: Average accuracy (%) and standard deviation across models for each task category under different evaluation setups.

These findings demonstrate that **high-fidelity visual input is essential for activating geometric reasoning** in current VLMs. While **sparse or dotted stimuli may suffice for simple linear tasks, they fall short for complex shape recognition and spatial inference**, underscoring the need for more expressive and grounded visual processing in geometric reasoning benchmarks.

Not all VLMs leverage full images equally. Figure 2 shows the gain from dots to full images. InternVL-2.5-1B and Qwen2-VL-Instruct improve consistently, whereas InternVL-3, G-LLaVA-13, and URSA gain little or regress, signaling weak visual grounding or instruction-following most pronounced on numeric items.

This phenomenon may correlate with several factors: poor handling of large or complex images, degraded adherence to structured prompts in multimodal settings, or overfitting to irrelevant visual patterns that misalign with the task objective. These cases highlight a critical limitation – larger image context can confuse undertrained or improperly aligned VLMs, leading to performance drops.

The takeaway message here is that **merely accessing visual data is not enough; it's essential to effectively ground and integrate image features to fully capitalize on the advantages offered by complete image input**.

Task sensitivity to modality. Some tasks (e.g., remarkable triangle lines, simple circle properties) show dramatic performance gains when moving from text-only to full visual input (e.g., Qwen2.5: from 0.0% on text-only task to 99.0% on full image task on unstable simple circle properties with VL model version). Others (e.g., symmetry) remain relatively stable, indicating that some tasks are more sensitive to modality than others. This uncovers another pattern: **benchmarking across modality types reveals where geometry is textually recoverable versus inherently visual.**

Instruction-following abilities degradation risk in math-specialized models. Math-specialized text-only models (e.g., Qwen2.5-Math-7B-Instruct) often ignore structured prompts and misformat outputs, performing worse on classification and unstable tasks than general instruction-tuned peers. This suggests that domain-specific fine-tuning can erode broad instruction adherence by overfitting to rigid mathematical formats.

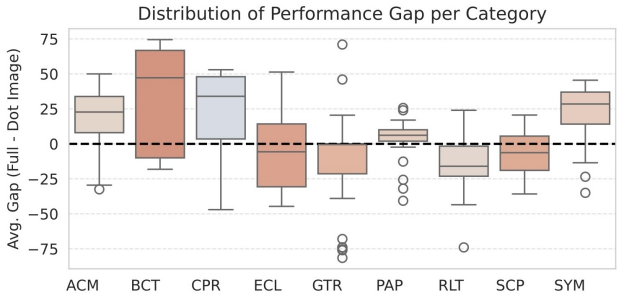

Figure 2: Distribution of model-level performance gaps per task category, comparing 'text with full image' to 'text with dotted image' setups.

Additionally, we note the high standard deviations across task categories (see Table 3), which likely reflect the varying difficulty of tasks within each category. Some tasks are straightforward, while others require an understanding of complex geometry and precise calculation of answers, resulting in uneven performance across models. Differences in model capabilities and training objectives also contribute to this variability.

Humans breeze through dotted-image multiple-choice items, yet struggle with numeric perimeter/area estimates – sometimes scoring below InternVL, Phi-3.5, and LLaVA-Mini on ECL tasks.

Vision Encoders Linear Probing

To measure how much of each geometry task is already linearly separable in contemporary vision embeddings (Alain and Bengio 2017), we carry out 2 stages of linear probing.

Estimating task solvability in standalone vision features. We extract image embeddings with the OpenAI CLIP-ViT-B/32 encoder (Radford et al. 2021) and train a linear classifier on them. For every one of eight binary geometry tasks – collinearity, diameter, semicircle-triangle, parallelism, perpendicular lines, right angles, shape symmetry, and symmetry points – we generate 10K training and 10K test images. In the multi-task regime this amounts to 80K training samples.

We probe each VLM's vision backbone and CLIP to as-

Tran DS	Config	All_F	All_D	C-BCT-CT_F	C-BCT-CT_D	U-CPR-CDT_F	U-CPR-CDT_D	U-SCP-STT_F	U-SCP-STT_D	U-PAP-PLT_F	U-PAP-PLT_D	U-PAP-PT_F	U-PAP-PT_D	U-PAP-RAT_F	U-PAP-RAT_D	C-SYM-SST_F	C-SYM-SST_D	U-SYM-ST_F	U-SYM-ST_D
All	UF	97.4	58.6	98.0	50.0	100.0	52.0	100.0	48.0	94.0	77.0	93.0	61.0	100.0	68.0	100.0	67.0	94.0	46.0
	FF	92.6	61.4	98.0	57.0	100.0	54.0	96.0	54.0	85.0	80.0	68.0	66.0	100.0	78.0	100.0	49.0	94.0	53.0
	UD	56.6	83.9	47.0	93.0	79.0	91.0	41.0	61.0	68.0	91.0	56.0	73.0	67.0	93.0	38.0	96.0	57.0	73.0
	FD	49.2	72.9	47.0	92.0	51.0	66.0	48.0	59.0	59.0	84.0	43.0	61.0	49.0	84.0	38.0	74.0	59.0	63.0
Separate	UF	97.4	58.6	100.0	63.0	100.0	52.0	100.0	49.0	93.0	87.0	95.0	46.0	100.0	55.0	100.0	52.0	94.0	51.0
	FF	92.6	61.4	99.0	48.0	100.0	52.0	100.0	51.0	88.0	83.0	70.0	58.0	100.0	55.0	100.0	52.0	91.0	49.0
	UD	56.6	83.9	49.0	95.0	50.0	97.0	51.0	70.0	81.0	93.0	69.0	75.0	68.0	93.0	60.0	99.0	52.0	67.0
	FD	49.2	72.9	47.0	82.0	47.0	67.0	49.0	61.0	71.0	87.0	58.0	68.0	77.0	90.0	41.0	89.0	60.0	65.0

Table 4: Linear-probe accuracy (binary) for models trained jointly (top 4 rows) or per task (bottom 4). Configs: UF/FF = unfrozen/frozen encoder on full images; UD/FD = unfrozen/frozen on dot images. Evaluation tags: BC = basic-coordinate, CP = circle-properties, PP = parallelism-perpendicularity, Sym = symmetry; subscript F/D marks full vs. dot test images.

sess (i) tasks' linear separability and (ii) geometric improvements from vision-language pre-training. Four setups are tested: frozen/fine-tuned encoders with full/dot-only images, using linear heads pooling all patch tokens. Training includes task-specific probes (10K samples each), multi-task probes (80K images), and cross-task transfer tests.

Probing Results Analysis

Full vs. dotted diagrams. A fine-tuned ViT-B/32 linear probe achieves 97% accuracy on full images but drops to 58% on dot-only tests, showing CLIP's reliance on global shape cues. Retraining on dots reverses this pattern: 85% on dots, 57% on full images.

Effect of freezing. Freezing the backbone costs roughly five points on full images (FF = 92%) and yields the weakest dot performance (FD $\sim$ 73%), indicating that a small amount of adaptation is important for geometry.

Cross-task transfer (Table 5). Parallel-line probes push right-angle accuracy to 86%, and symmetry probes lift circle-symmetry to 81%, confirming family-level transfer. Still, VLMs lag on NoReGeo, implying language training overlooks these geometric cues.

Train Dataset	Eval Dataset	Configs with Accuracy
U-SYM-ST	C-SYM-SST	UD (81.0)
U-SYM-ST	U-CPR-CDT	UD (74.0)
C-SYM-SST	U-CPR-CDT	FF (79.0)
C-SYM-SST	C-BCT-CT	FD (86.0)
U-PAP-PLT	U-PAP-PT	UD (75.0), UD (74.0), UF (71.0)
U-PAP-PLT	C-BCT-CT	FD(90.0), FF (78.0)
U-PAP-PT	C-SYM-SST	FF (72.0)
U-PAP-PT	U-PAP-PLT	UD (90.0), UD (86.0), FD (85.0)
U-PAP-PT	U-PAP-RAT	UD (80.0), FD (78.0), UF (71.0)
U-PAP-PT	U-CPR-CDT	UD (83.0), FF (76.0), UF (75.0)
U-SCP-STT	C-SYM-SST	FF (90.0)
U-CPR-CDT	U-SCP-STT	UF (80.0), FF (77.0)
C-BCT-CT	U-SCP-STT	UF (95.0)
C-BCT-CT	U-CPR-CDT	UF (100.0), FF (73.0)

Table 5: Linear-probing training transfer.

Conclusion

We introduced NoReGeo, a cross-modal benchmark of elementary geometry problems designed to assess whether LLMs and VLMs can answer spatial questions *without* relying on explicit reasoning. Across more than 45 state-of-the-art models, we find that most struggle with tasks that humans solve through immediate geometric intuition, often producing unnecessary chain-of-thought explanations even when instructed otherwise. In contrast, a frozen vision encoder paired with a simple linear probe performs almost perfectly, indicating that the essential geometric cues are already embedded in visual representations.

These findings position NoReGeo as a practical tool for probing latent geometric competence in modern foundation models and for selecting models when fast, geometry-aware inference is required. Looking ahead, we plan to explore how fine-tuning and representation alignment influence generalization across geometric concepts and whether models can be encouraged to develop more robust, human-like geometric intuition.

Acknowledgments

Innopolis University authors were supported by the Research Center of the Artificial Intelligence Institute at Innopolis University. Financial support was provided by the Ministry of Economic Development of the Russian Federation (No. 25-139-66879-1-0003).

References

Abraham, S. S.; Alirezaie, M.; and Raedt, L. D. 2024. CLEVR-POC: Reasoning-Intensive Visual Question Answering in Partially Observable Environments. arXiv:2403.03203.

Alain, G.; and Bengio, Y. 2017. Understanding intermediate layers using linear classifier probes.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Wang, J.; Qiao, Y.; Lin, D.; and Zhao, F. 2024. Are We on the Right Way for Evaluating Large Vision-Language Models? arXiv:2403.20330.

Chevalier-Boisvert, M.; Bahdanau, D.; Lahlou, S.; Willems, L.; Saharia, C.; Nguyen, T. H.; and Bengio, Y. 2019. BabyAI: A Platform to Study the Sample Efficiency of Grounded Language Learning. arXiv:1810.08272.

Dong, Y.; Ruan, C. F.; Cai, Y.; Lai, R.; Xu, Z.; Zhao, Y.; and Chen, T. 2024. Xgrammar: Flexible and efficient structured generation engine for large language models. *arXiv preprint arXiv:2411.15100*.

Gao, J.; Pi, R.; Zhang, J.; Ye, J.; Zhong, W.; Wang, Y.; Hong, L.; Han, J.; Xu, H.; Li, Z.; et al. 2023. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hong, H.; Qiao, Y.; Wang, S.; Liu, J.; and Wu, Q. 2025. General Scene Adaptation for Vision-and-Language Navigation. arXiv:2501.17403.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. arXiv:2310.06825.

Kazemi, M.; Alvari, H.; Anand, A.; Wu, J.; Chen, X.; and Soricut, R. 2024. GeomVerse: A Systematic Evaluation of Large Models for Geometric Reasoning. In *AI for Math Workshop @ ICML 2024*.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J. E.; Zhang, H.; and Stoica, I. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. arXiv:2309.06180.

Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.-W.; Galley, M.; and Gao, J. 2024. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. arXiv:2310.02255.

Lu, P.; Gong, R.; Jiang, S.; Qiu, L.; Huang, S.; Liang, X.; and Zhu, S.-C. 2021. Inter-GPS: Interpretable Geometry Problem Solving with Formal Language and Symbolic Reasoning. In *The 59th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Luo, R.; Zheng, Z.; Wang, Y.; Ni, X.; Lin, Z.; Jiang, S.; Yu, Y.; Shi, C.; Chu, R.; Zeng, J.; et al. 2025a. Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. *arXiv preprint arXiv:2501.04686*.

Luo, S.; Zhu, Z.; Yuan, Y.; Yang, Y.; Shan, L.; and Wu, Y. 2025b. GeoGramBench: Benchmarking the Geometric Program Reasoning in Modern LLMs. arXiv:2505.17653.

Ma, W.; Chen, H.; Zhang, G.; Chou, Y.-C.; de Melo, C. M.; and Yuille, A. 2025. 3DSRBench: A Comprehensive 3D Spatial Reasoning Benchmark. arXiv:2412.07825.

National Council of Teachers of Mathematics. 2000. *Principles and Standards for School Mathematics*. Reston, VA: National Council of Teachers of Mathematics. ISBN 0-87353-480-8.

National Governors Association Center for Best Practices and Council of Chief State School Officers. 2010. Common Core State Standards for Mathematics. Accessed: 2025.

Peng, S.; Fu, D.; Gao, L.; Zhong, X.; Fu, H.; and Tang, Z. 2024. Multimath: Bridging visual and mathematical reasoning for large language models. *arXiv preprint arXiv:2409.00147*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Rodionov, F.; Eldesokey, A.; Birsak, M.; Femiani, J.; Ghanem, B.; and Wonka, P. 2025. PlanQA: A Benchmark for Spatial Reasoning in LLMs using Structured Representations. arXiv:2507.07644.

Shi, W.; Hu, Z.; Bin, Y.; Liu, J.; Yang, Y.; Ng, S.-K.; Bing, L.; and Lee, R. K.-W. 2024. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*.

Shridhar, M.; Thomason, J.; Gordon, D.; Bisk, Y.; Han, W.; Mottaghi, R.; Zettlemoyer, L.; and Fox, D. 2020. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. arXiv:1912.01734.

Team, Q. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Trinh, T. H.; Wu, Y.; Le, Q. V.; He, H.; and Luong, T. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995): 476–482.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, X.; Wang, Y.; Zhu, W.; and Wang, R. 2025. Do Large Language Models Truly Understand Geometric Structures? In *The Thirteenth International Conference on Learning Representations*.

Willard, B. T.; and Louf, R. 2023. Efficient Guided Generation for LLMs. *arXiv preprint arXiv:2307.09702*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, K.; Russakovsky, O.; and Deng, J. 2019. SpatialSense: An Adversarially Crowdsourced Benchmark for Spatial Relation Recognition. arXiv:1908.02660.

Zhang, J.; Li, Z.-Z.; Zhang, M.-L.; Yin, F.; Liu, C.-L.; and Moshfeghi, Y. 2024a. GeoEval: Benchmark for Evaluating LLMs and Multi-Modal Models on Geometry Problem-Solving. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 1258–1276. Bangkok, Thailand: Association for Computational Linguistics.

Zhang, M.-L.; Yin, F.; and Liu, C.-L. 2023. A multi-modal neural geometric solver with textual clauses parsed from diagram. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23. ISBN 978-1-956792-03-4.

Zhang, R.; Jiang, D.; Zhang, Y.; Lin, H.; Guo, Z.; Qiu, P.; Zhou, A.; Lu, P.; Chang, K.-W.; Gao, P.; and Li, H. 2024b. MathVerse: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems? arXiv:2403.14624.