TangramPuzzle: Evaluating Multimodal Large Language Models with Compositional Spatial Reasoning

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) have achieved remarkable progress in visual recognition and semantic understanding. Nevertheless, their ability to perform precise compositional spatial reasoning remains largely unexplored. Existing benchmarks often involve relatively simple tasks and rely on semantic approximations or coarse relative positioning, while their evaluation metrics are typically limited and lack rigorous mathematical formulations. To bridge this gap, we introduce **TangramPuzzle** [1], a geometry-grounded benchmark designed to evaluate compositional spatial reasoning through the lens of the classic Tangram game. We propose the Tangram Construction Expression (TCE), a symbolic geometric framework that grounds tangram assemblies in exact, machine-verifiable coordinate specifications, to mitigate the ambiguity of visual approximation. We design two complementary tasks: Outline Prediction, which demands inferring global shapes from local components, and End-to-End Code Generation, which requires solving inverse geometric assembly problems. We conduct extensive evaluation experiments on advanced open-source and proprietary models, revealing an interesting insight: MLLMs tend to prioritize matching the target silhouette while neglecting geometric constraints, leading to distortions or deformations of the pieces.

1 Introduction

Multimodal Large Language Models (MLLMs) have achieved strong performance on a broad range of vision–language tasks, including visual understanding (Wang et al., 2025c; Fang et al., 2025; Ding et al., 2025), document question answering (Wu et al., 2025b; Huang et al., 2025a), OCR (Chen et al., 2025a; Huang et al., 2025b; Greif et al., 2025), and chart interpretation (Chen et al., 2025b; Zhao et al., 2025). Recently, both training and evaluation have been increasingly shifting from semantic recognition toward more demanding tasks that require precise spatial understanding and reasoning, such as visual grounding (Tang et al., 2025b; Xu et al., 2025d; Ma et al., 2025b), GUI agent (Cheng et al., 2025b; Xiao et al., 2025), and visual reasoning in complex scenes (Kuang et al., 2025; Li et al., 2024). Spatial reasoning is not merely about recognizing objects, but about constructing and validating structured geometric relationships that support reliable decision making.

Despite rapid progress, existing spatial reasoning tasks for MLLMs still exhibit several limitations. First, many tasks focus on coarse relational concepts (e.g., left/right, front/behind, adjacency) (Liu

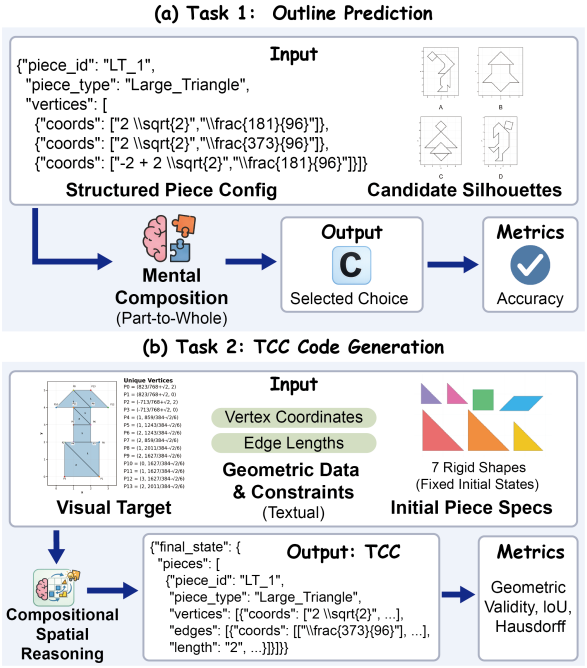

Figure 1: Illustration of the two spatial reasoning tasks. (a) Task 1 global shape inference; (b) Task 2 constrained shape decomposition.

[1]We will open-source all data and code after the anonymity review process ends.

et al., 2025; Deng et al., 2025; Wu et al., 2025a), which do not fully capture reasoning under strict geometric and physical constraints. Second, spatial problems are often specified using natural language descriptions or multiple-choice formats without a formal geometric representation (Rodionov et al., 2025; Ma et al., 2025a), relying more on linguistic reasoning ability than visual reasoning. In addition, evaluation commonly relies on text answer matching or human/LLM judgment, which is particularly problematic for constructive tasks where multiple valid solutions may exist. ***These issues make it difficult to precisely measure whether a model truly "compositionally reasons" in space or merely exploits superficial cues.***

In this work, we are inspired by a classic but underexplored task of compositional spatial reasoning: tangram puzzles (Zhao et al., 2022; Bofferding and Aqazade, 2023; Bofferding and Zhu, 2023). A tangram requires assembling seven fixed rigid polygonal pieces into a target silhouette through Euclidean transformations (translation, rotation, and optionally reflection), while satisfying non-overlap, full coverage, and topological validity constraints. This setting naturally demands multi-step compositional spatial reasoning from local parts to a global shape. These properties make tangrams a compact yet expressive testbed for complex spatial reasoning beyond relative positioning.

However, prior tangram-related resources and studies only focus on overall interpretation of shapes or perceptual approximations (Zhang et al., 2024; Ji et al., 2022), and typically lack (i) a standardized, machine-verifiable geometric representation, and (ii) a strict evaluation protocol that validates rigidity, physical feasibility, and solution correctness. As a result, task difficulty is undercontrolled, and evaluation accuracy remains limited.

To address these gaps, we introduce TangramPuzzle, a benchmark designed to evaluate compositional spatial reasoning in MLLMs with rigorous geometric constraints. We propose the Tangram Construction Expression (TCE), a symbolic geometry schema that represents each instance with exact algebraic coordinates and structural relations, strictly eliminating floating-point ambiguity. Built upon TCE, the benchmark includes two complementary tasks, as shown in Figure 1: (1) Outline Prediction, where models infer the global silhouette from a given final arrangement of pieces; and (2) End-to-End Tangram Solution Generation, where models construct a complete assembly that exactly fills a given target outline. We evaluate outputs with a constraint-based verifier that checks syntactic validity, rigid-shape preservation, and physical feasibility (e.g., non-overlap and connectivity), and then measures silhouette fidelity using IoU and Hausdorff distance, avoiding ill-posed comparisons under multi-solution ambiguity. TangramPuzzle is constructed through a multi-stage pipeline involving source filtering, an interactive annotation tool with snapping, symbolic normalization into exact expressions, and human-in-the-loop validation. Extensive experiments across a wide range of MLLMs show that TangramPuzzle remains highly challenging, revealing systematic failure modes in both compositional outline reasoning and constraint-satisfying solution generation. Our main contributions are as follows:

- We introduce TangramPuzzle, a tangram-inspired benchmark for compositional spatial reasoning in MLLMs, with a rigorous and machine-verifiable math representation.
- We design the data construction pipeline for two tasks that jointly evaluate discriminative global-shape inference and constructive inverse assembly generation.
- We conduct comprehensive evaluation of advanced MLLMs, highlighting persistent limitations in complex spatial reasoning.

2 Related Work

2.1 General Multimodal Benchmarks

Standard benchmarks for Multimodal Large Language Models (MLLMs) have transitioned from foundational tasks like visual question answering (VQA) (Zhong et al., 2025; Liu et al., 2024; Chen et al., 2024a; Cheng et al., 2025c) and image captioning (Dong et al., 2024; Cheng et al., 2025a; Lu et al., 2025) to comprehensive assessments of emergent capabilities. Recent benchmarks target diverse competencies, including reasoning ability (Guo et al., 2025; Yuan et al., 2025; Yue et al., 2025; Xiao et al., 2024), where challenging datasets like Humanity's Last Exam (Phan et al., 2025) and ZeroBench (Roberts et al., 2025) test the upper bounds of expert-level cognition; document and chart understanding (Wang et al., 2024; Fu et al., 2024a; Xia et al., 2025; Ouyang et al., 2025; Masry et al., 2025); as well as visual grounding (Paiss et al., 2023; Brazil et al., 2023), hallucination evaluation (Li et al., 2023; Guan et al., 2024; Leng et al.,

2024; Chen et al., 2024b; Yang et al., 2025b), and multi-image comprehension (Cheng et al., 2025d; Fu et al., 2024b; Jiang et al., 2024; Wang et al., 2025a). Furthermore, the assessment of agentic capabilities (Li et al., 2025; Xie et al., 2025; Rawles et al., 2025; Zheng et al., 2025a) centers on GUI interpretation and precise element interaction.

2.2 Spatial Reasoning Benchmarks

Spatial reasoning is essential for MLLMs to perceive and interact with the physical world (Zheng et al., 2025b). Early research emphasized spatial understanding, as SpatialBench (Xu et al., 2025b) and RealWorldQA (xAI, 2024) evaluate relative positioning and depth in natural images. As research advanced, OmniSpatial (Jia et al., 2025) and MMSI-Bench (Yang et al., 2025a) extended the scope to dynamic and multi-view scenarios, assessing complex spatial interactions and multi-image reasoning relationships. LEGO-Puzzles (Tang et al., 2025a) and ORIGAMISPACE (Xu et al., 2025c) emphasize multi-step spatial reasoning, utilizing block assembly and origami tasks to probe physical constraints and geometric transformations. GeoSense (Xu et al., 2025a) and SolidGeo (Wang et al., 2025b) require applying strict geometric principles to solve plane and solid geometry tasks. In the Tangram domain, Tangram (Zhang et al., 2024) targets basic element recognition and counting; KILOGRAM (Ji et al., 2022) interprets shapes as semantic concepts rather than constructive geometric logic; and TANGAN (Yamada et al., 2025) acts as a specialized solver treating assembly as an optimization problem, rather than evaluating MLLMs.

Unlike previous approaches that rely on perceptual approximations, our benchmark distinguishes through complex spatial reasoning, rigorous mathematical expression, and strict constraint evaluation, ensuring the exact geometric validity of solutions.

3 Method

3.1 Task Definition

The Tangram is a classic dissection puzzle requiring the arrangement of seven rigid polygons to form specific shapes, serving as an ideal testbed for compositional spatial reasoning. While natural language is expressive, relying solely on verbal descriptions for geometric assembly leads to substantial ambiguity and imprecision, making reliable evaluation difficult due to the lack of spatial granularity. In contrast, TangramPuzzle is governed by strict Euclidean geometry and topological constraints, motivating a formal, coordinate-based representation that grounds visual inputs in precise mathematical formulations. This design enables unambiguous task definition, automatic verification, and fine-grained evaluation of spatial reasoning abilities. TangramPuzzle consists of two challenging tasks designed to evaluate the spatial reasoning capabilities of MLLMs, ranging from passive spatial perception to active solution generation.

Geometry Description by TCE Formula To represent tangram configurations in a precise and machine-verifiable manner, we introduce the Tangram Construction Expression (TCE), a symbolic geometry description language for 2D tangram assembly. Each TCE instance consists of five core components: instance_id, target_outline, initial_state, final_state, and adjacency_graph.

The field **instance_id** uniquely identifies each puzzle instance and serves as a stable key for indexing, retrieval, and evaluation. Both **initial_state** and **final_state** describe the geometric states of the seven tangram pieces before and after assembly. In TCE, each piece is specified by its piece type, a set of vertex coordinates, explicit edge relations, and a center point. In addition, final_state includes a transform_matrix that parameterizes the rigid motion applied to each piece to reach the assembled configuration, including translation, rotation angle, and optional reflection. Because tangram geometries frequently involve irrational numbers, finite decimal representations induce precision loss. To ensure geometric rigor, TCE encodes all quantities as exact algebraic expressions in LaTeX. The **target_outline** specifies the global silhouette produced by the assembled tangram. It is represented by the outline's vertex coordinates together with its edge relations. The **adjacency_graph** captures local interactions between pieces in the final configuration, indicating which pieces are adjacent. The representation of TCE is shown in Figure 2.

$$
\begin{aligned}
\textbf{TCE} &= \langle \texttt{id}, \Omega_{target}, \Sigma_{init}, \Sigma_{final}, G_{adj} \rangle \\
\texttt{id} &= \texttt{"swan\_cfg\_01"} \\
\Omega_{target} &= \{V : [(0,0), (4, 2\sqrt{2}), ...], E : [(v_0, v_1), ...]\} \\
\Sigma_{final} &= \{p_{T1} : \langle \mathbf{M} = \begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 2 \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}, V_{world} : [...] \rangle, ..\} \\
G_{adj} &= \{(T1, S1), (S1, T3), ...\}
\end{aligned}
$$

Figure 2: Compact representation of a TCE instance.

Task 1: Outline Prediction This setting tests whether MLLMs can synthesize a global visual shape from local component arrangements. Rather than performing conventional visual recognition, the model is required to mentally compose the union of seven discrete tangram pieces and reason about their resulting overall shape. The input consists of the textual TCE specification of the seven tangram pieces (including piece types and exact vertex coordinates) and a candidate image displaying four distinct silhouette options (A–D). The model is required to select the correct matching silhouette from the candidate set. Accuracy and Invalid Rate are used as the evaluation metrics.

Task 2: End-to-End Tangram Solution Generation This task assesses the model's ability to solve the tangram puzzle under strict geometric constraints and generate a precise executable solution. It represents a complex inverse spatial reasoning problem where the model must decompose a given target silhouette into the seven fixed rigid shapes. The input consists of a visual image of the target outline with vertex coordinates explicitly annotated within the figure. To precisely define the boundary conditions, the model is also provided with textual geometric data, including the vertex coordinates of the target outline and its edge relationships, specifying which vertices connect to form edges and their respective lengths. Furthermore, the input includes the initial state specifications for the seven tangram pieces, detailing their types, initial vertex coordinates, and constituent edges. The model is required to output the complete Tangram Configuration Expression (TCE) in JSON format. This code must explicitly define the final state of all seven pieces, including their types, vertex coordinates, and edge relationships, ensuring they perfectly fill the target outline. The evaluation checks both the geometric validity (satisfaction of rigid body and non-overlap constraints) and the shape fidelity of the generated solution, measured by Intersection over Union (IoU) and Hausdorff Distance (d_H) against the ground truth.

3.2 Data Construction

To ensure high-quality geometric data, we implemented a multi-stage construction pipeline, as illustrated in Figure 3, comprising source filtering, interactive annotation, and symbolic refinement.

3.2.1 Raw Data Collection

We utilize the visual patterns from the KiloGram dataset as our foundational source. While KiloGram (Ji et al., 2022) provides a rich variety of tangram shapes, not all adhere to the strict topological constraints required for mathematical verification. Consequently, we perform a manual filtering process to exclude invalid configurations. Specifically, we remove patterns containing holes, disconnected pieces, or incomplete assemblies. Only configurations that form a single connected shape without internal voids are retained for further annotation.

3.2.2 Annotation Tool Development

Commonly used geometric software has proven insufficient for efficient and precise tangram annotation. Tools like GeoGebra[2] require users to compute coordinates before drawing and offer limited support for interactive manipulation, while manipulation-friendly platforms like Polypad[3] lack the functionality to export precise vertex coordinates. To address these limitations, we developed a bespoke web-based tangram annotation interface. To reduce human-induced alignment errors, we further incorporate vertex snapping and edge snapping mechanisms that automatically align coincident geometric features during assembly. The tool records the coordinates of all vertices in real time and exports complete annotations directly in a structured jsonl format, substantially accelerating the labeling process. Detailed human annotation procedures and the tool are provided in Appendix A.

3.2.3 Geometric Normalization

The raw annotations produced by the tool are stored in floating-point form, which is insufficient for exact geometric reasoning. To eliminate ambiguity caused by finite-precision decimals, we perform a post-processing step to normalize and convert all geometric quantities into exact symbolic representations. Specifically, each assembled configuration is translated such that its lower boundary lies on the x-axis and its left boundary lies on the y-axis, yielding a canonical placement. We then apply deterministic conversion rules to transform all vertex coordinates into precise algebraic expressions, represented in LaTeX format. Based on these coordinates, we further extract the global outer contour, edge relations, and piece adjacency, and serialize the result into the unified TCE representation.

[2] https://www.geogebra.org/classic
[3] https://polypad.amplify.com/p#tangram

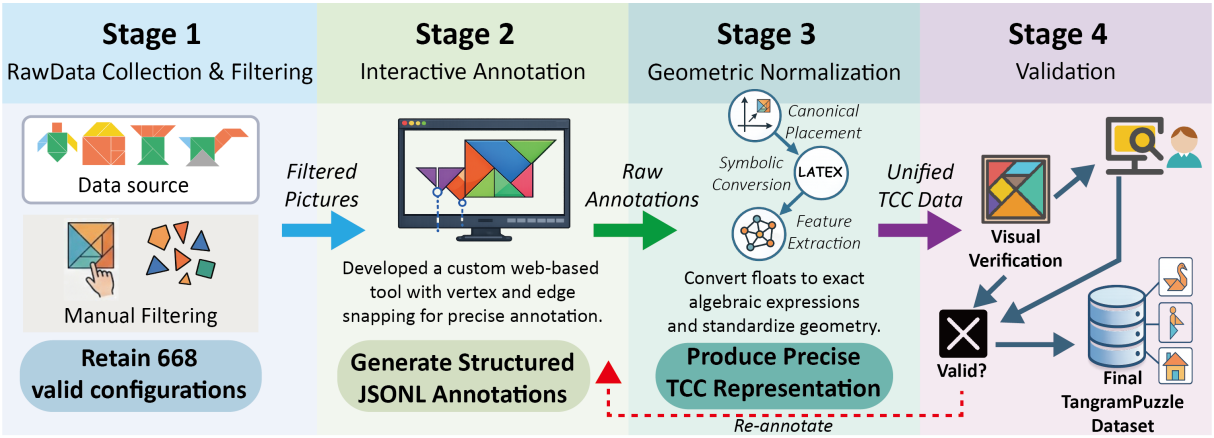

Figure 3: Overview of the TangramPuzzle data construction pipeline.

To generate samples for the Outline Prediction task, we leverage the geometric properties of the TCE. For each tangram instance, we first reconstruct the ground-truth silhouette accurately based on the corresponding information encoded in the TCE. To construct the multiple-choice options, we randomly sample three distinct silhouettes from the remaining dataset as negative distractors. These distractors are combined with the ground-truth silhouette, randomly shuffled, and rendered into a single candidate image labeled A through D. This automated pipeline ensures that the ground-truth silhouette is geometrically consistent with the coordinate-based input while maintaining visual diversity among the options.

3.2.4 Human Filtering

To ensure data correctness, we perform careful manual verification through visualization. Each annotated instance is rendered to display the assembled configuration, highlighting individual pieces and their boundaries. Instances exhibiting misaligned vertices, unintended gaps between adjacent edges, or overlaps between pieces are promptly flagged as invalid. Such cases are re-annotated using the tool, followed by repeated verification and conversion. This iterative process continues until all retained instances satisfy the geometric validity constraints required by TCE.

3.2.5 Data Statistics

TangramPuzzle exhibits both rich structural diversity and strict geometric rigor. The target silhouettes cover a wide range of semantic categories, including abstract shapes, everyday objects, and animals, spanning a broad and challenging spectrum of difficulty from simple, compact forms to complex, articulated silhouettes with intricate boundaries. In total, the dataset contains 668 unique tangram configurations. Each configuration is used to instantiate both Task 1 and Task 2, resulting in 1,336 problem instances overall.

3.3 Evaluation Metrics

Task 1: Outline Prediction To evaluate vision-grounded spatial perception, we measure the standard classification **Accuracy (Acc)**, which calculates the percentage of instances where the model selects the correct target outline. Additionally, to assess the model's instruction-following capability and robustness, we report the **Invalid Rate**, defined as the proportion of responses that fail to map to any valid candidate option. Invalid outputs may arise from formatting errors or from generating content unrelated to the candidate options, reflecting common failure modes in instruction execution, in which case no interpretable decision can be extracted. Detailed metric computations for Task 1 are provided in Appendix B.1.

Task 2: End-to-End Solution Generation Evaluating assembly generation is non-trivial because a single target silhouette may admit multiple valid internal arrangements. Direct comparison with a single ground truth is therefore ill-posed. Instead, we propose a hierarchical Constraint-based Evaluation Framework that validates predictions in two stages. Detailed metric definitions and computation procedures for Task 2 are provided in Appendix B.2. Concretely, the framework comprises:

- **Stage 1: Constraint Validation.** We first check whether the generated assembly satisfies the required physical and geometric constraints by identifying three types of errors:

Syntax Error (TSE) checks whether the output conforms to the TCE format and piece count; **Rigid Geometry Error (RGE)** verifies that the area and perimeter of each piece remain unchanged to prevent shape distortion; and **Physical Error (PE)** detects if any two pieces exhibit impermissible overlap or if the global union of all pieces fails to form a single connected component. The metric **Validation Pass Rate (VPR)** reports the percentage of samples that satisfy all these constraints.

- **Stage 2: Shape Similarity.** We further quantify how well the constructed shape matches the target silhouette using two metrics: **Intersection over Union (IoU)** measures the global area overlap between the predicted assembly and the target, while **Hausdorff Distance** (d_H) captures fine-grained boundary deviations to penalize shape outliers.

4 Experiment

4.1 Experimental Setting

Baselines We evaluate TangramPuzzle across a diverse set of models. For open-source models, we evaluate Qwen3-VL-8B-Instruct, Qwen3-VL-32B-Instruct (Bai et al., 2025), InternVL3-78B (Zhu et al., 2025), DeepSeek-OCR (Wei et al., 2025), DeepSeek-VL2 (Wu et al., 2024), and GLM-4.6V (Z.ai, 2025). For commercial models, we include GPT-5.2 (OpenAI, 2025), Gemini3-Pro (Google DeepMind, 2025), and Claude-Sonnet-4.5 (Anthropic, 2025). We utilize the official chat templates and default generation parameters for all models to ensure a fair comparison.

Details All experiments are conducted by calling the corresponding model APIs. We run the evaluation pipeline on a workstation equipped with an NVIDIA RTX 2080 Ti GPU, and use a unified implementation to send requests, parse outputs, and compute all metrics across models.

4.2 Results

The results for Task 1 are summarized in Table 1, revealing substantial performance differences across models. This task assesses the model's ability to mentally compose a silhouette from disjointed parts and identify the correct visual match accurately. Among open-source baselines, Qwen3-VL-32B-Instruct stands out with an accuracy of 73.05%, approaching the performance of the closed-source model GPT-5.2, while smaller or earlier architectures struggle to exceed 40% accuracy. For commercial models, Gemini3-Pro achieves near-ceiling performance with an accuracy of 98.65%, underscoring its superior capacity for fine-grained visual perception and holistic mental shape composition regarding global silhouettes. For most models, the rate of invalid outputs remains low, suggesting a strong ability to follow instructions and produce interpretable decisions. However, DeepSeek-OCR and DeepSeek-VL2 exhibit noticeably higher invalid rates, suggesting that under the stress of this complex spatial reasoning task, these models often fail to form a decisive conclusion or collapse into generating irrelevant content.

Table 1: Performance comparison on Task 1.

Model	Invalid (↓)	Acc (↑)
Qwen3-VL-8B (Bai et al., 2025)	**0**	37.57
Qwen3-VL-32B (Bai et al., 2025)	0.15	73.05
InternVL3-78B (Zhu et al., 2025)	0.75	29.19
DeepSeek-OCR (Wei et al., 2025)	12.87	22.46
DeepSeek-VL2 (Wu et al., 2024)	10.18	25.45
GLM-4.6V (Z.ai, 2025)	**0**	30.84
GPT-5.2 (OpenAI, 2025)	**0**	77.54
Gemini3-Pro (Google DeepMind, 2025)	**0**	**98.65**
Claude-Sonnet-4.5 (Anthropic, 2025)	0.45	51.20

Table 2 presents the quantitative results for Task 2. Although most models demonstrate proficiency in maintaining valid JSON syntax, as reflected by relatively low TSE scores, constraint-related failures remain widespread. In particular, Rigid Geometry Errors (RGE) and Physics Errors (PE) are highly prevalent, indicating that models struggle to comply with the fundamental geometric rules of the puzzle. One notable observation is the pronounced discrepancy between silhouette quality and the Success rate. Top-tier models like Claude-Sonnet-4.5 and GPT-5.2 achieve high IoU scores yet fail to produce a single valid solution (0% Success). Further inspection reveals that the accompanying high RGE and PE scores indicate that this apparent visual fidelity is achieved through "cheating," namely by impermissibly distorting rigid pieces or overlapping them to force a visual match. As a result, the vast majority of evaluated models fail to produce even a single geometrically valid solution. By contrast, Gemini3-Pro stands out as the only model demonstrating robust geometric reasoning, achieving the highest validation pass rate and success score while also maintaining superior silhouette quality.

Table 2: Quantitative results for the End-to-End TCE Generation task (Task 2).

Model	Constraint error (↓)			VPR (↑)	Silhouette Quality		Success (↑)
	TSE (↓)	RGE (↓)	PE (↓)		IoU (↑)	Hausdorff (↓)	
Qwen3-VL-8B-Instruct (Bai et al., 2025)	23.35	40.42	80.69	0	20.25	7.8138	0
Qwen3-VL-32B-Instruct (Bai et al., 2025)	7.93	83.08	94.46	0	50.55	1.3949	0
InternVL3-78B (Zhu et al., 2025)	6.59	91.77	92.96	0	36.15	3.1756	0
GLM-4.6V (Z.ai, 2025)	6.29	91.02	93.56	0.15	43.79	1.3745	0.15
GPT-5.2 (OpenAI, 2025)	7.19	90.12	92.81	0	58.49	0.8375	0
Gemini3-Pro (Google DeepMind, 2025)	8.83	**37.57**	**65.57**	**22.60**	**85.93**	**0.3728**	**21.56**
Claude-Sonnet-4.5 (Anthropic, 2025)	**0.60**	98.20	97.90	0.15	61.61	0.5696	0

Table 3: Evaluation results on Task 2 under In-Context Learning and Visual-Centric settings.

Model	Constraint error (↓)			VPR (↑)	Silhouette Quality		Success (↑)
	TSE (↓)	RGE (↓)	PE (↓)		IoU (↑)	Hausdorff (↓)	
In-Context Learning							
Qwen3-VL-8B-Instruct (Bai et al., 2025)	33.98	57.04	66.17	0.15	28.48	3.5051	0
Qwen3-VL-32B-Instruct (Bai et al., 2025)	**6.74**	87.28	93.71	0	48.65	1.4885	0
GPT-5.2 (OpenAI, 2025)	9.43	86.98	90.72	0	60.30	0.7386	0
Gemini3-Pro (Google DeepMind, 2025)	10.78	**34.28**	**61.53**	**26.20**	**87.36**	**0.3494**	**24.85**
Visual-Centric setting							
Qwen3-VL-8B-Instruct (Bai et al., 2025)	15.42	17.07	86.23	0	11.39	12.2784	0
Qwen3-VL-32B-Instruct (Bai et al., 2025)	**9.58**	74.70	93.26	0.15	43.69	2.7822	0.15
DeepSeek-OCR (Wei et al., 2025)	100	**0**	**1.35**	0	11.72	7.8063	0
DeepSeek-VL2 (Wu et al., 2024)	99.85	0.15	13.32	0	12.08	5.4868	0
GPT-5.2 (OpenAI, 2025)	39.67	58.23	58.08	0	53.67	1.2140	0
Gemini3-Pro (Google DeepMind, 2025)	16.77	32.34	56.89	**25.00**	**82.75**	**0.4201**	**22.75**

4.3 Analysis

Effect of In-context Learning (ICL) To examine whether the bottleneck in Task 2 arises from difficulties in following the TCE schema or from deeper limitations in spatial reasoning, we apply in-context learning (ICL) by prepending three solved examples to the prompt, as shown in Table 3. Contrary to the expectation that examples stabilize output, we observe a trade-off: the introduction of dense symbolic contexts tends to increase syntactic errors (TSE), suggesting a cognitive load that distracts models from strict formatting requirements. However, for successfully parsed responses, ICL generally improves silhouette quality (IoU), indicating that while examples help refine shape approximation, they cannot spontaneously induce a fundamental understanding of rigid body constraints in models that lack them. A detailed analysis of in-context learning is provided in Appendix C.1.

Effect of Textual Geometry We further investigated the models' dependency on textual metadata through an ablation study in a Visual-Centric setting, as shown in Table 3. In this setup, we removed the textual description of the target outline (i.e., vertex and edge lists), providing only the outline image with annotated coordinates alongside the initial piece descriptions. The results show a significant performance drop for most MLLMs, with increased hallucinations and syntax errors. This confirms that current models largely rely on textual "crutches" rather than precisely extracting geometric coordinates from visual inputs. A notable exception is Gemini3-Pro, which maintained high performance, demonstrating superior capability in grounding geometric data directly from vision. Textual geometry analysis is detailed in Appendix C.2.

Human Performance We evaluate human performance on TangramPuzzle using three expert participants. As shown in Table 4, human solvers rarely violate geometric constraints such as piece shape preservation or piece count correctness. Consequently, human performance exhibits a near-binary pattern: solutions are either successfully completed or not solved at all, with few intermediate or partially correct outcomes. Participants report an average difficulty rating of 4.1/5 and an average solution time of 6.7 minutes per instance. A detailed analysis of human solving strategies, ranging from systematic triangle decomposition to iterative trial-and-error, is provided in Appendix D.

Table 4: Human performance on Task 2.

Human	Success (↑)	Complexity (↓)	Time usage (↓)
Human 1	72	4.1	**6.3**
Human 2	**100**	**3.8**	6.6
Human 3	46	4.3	7.1
Avg.	72.67	4.1	6.7

Table 5: Human evaluation of MLLM on Task 2.

Model	Expert	PSM (↑)	PPC (↑)	NMS (↑)
Gemini3-Pro	Expert 1	4.2	4.3	4.6
	Expert 2	4.5	4.6	4.1
	Expert 3	4.2	4.5	4.0
	Avg.	**4.3**	**4.5**	**4.2**
GPT-5.2	Expert 1	1.6	1.5	1.8
	Expert 2	1.2	1.1	1.1
	Expert 3	1.5	1.2	1.6
	Avg.	1.4	1.3	1.5

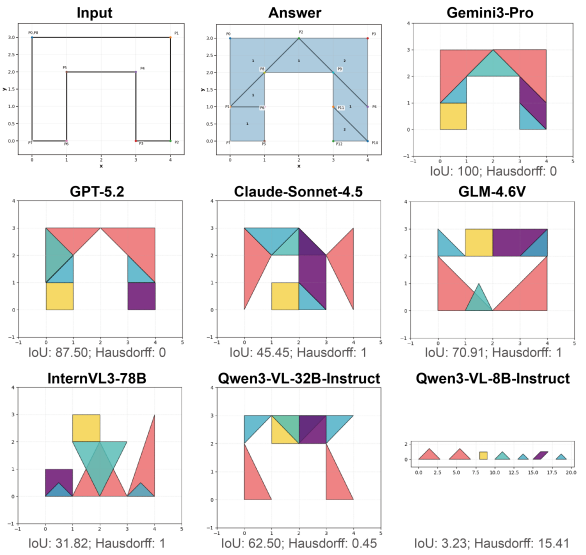

Figure 4: Visualization of a case from Task 2.

4.4 Human Evaluation on MLLMs

Automatic metrics primarily capture strict geometric validity, often overlooking perceptual plausibility. To address this, we conduct a human evaluation on Task 2 using three criteria rated on a 5-point scale: Perceptual Silhouette Match (PSM), which assesses the visual similarity of the assembly to the target; Piece Plausibility & Canonicality (PPC), which evaluates whether piece placements appear natural rather than visually strained; and Near-Miss Severity (NMS), which measures the proximity of an incorrect prediction to a successful solution. Table 5 summarizes the human evaluation results. Gemini3-Pro consistently receives high scores, indicating strong perceptual alignment with the target silhouette; even when automatic success is not achieved, many solutions are judged to be close to correct. In contrast, GPT-5.2 receives substantially lower scores, with frequent perceptual mismatches, implausible constructions, and failures that are judged to be far from correct.

4.5 Case Study

Figure 4 visualizes the generation results for a U-shaped target (Task 2). Gemini3-Pro achieves a perfect assembly, demonstrating its ability to effectively ground coordinate constraints to solve complex compositional problems. A critical failure mode observed across other models is the tendency to prioritize visual silhouette matching at the expense of geometric constraints. For instance, Claude-Sonnet-4.5 attempts to force an alignment with the target contour by non-rigidly elongating the edges of the left and right triangles, whereas InternVL3-78B generates severe overlaps to maximize area coverage. Meanwhile, GPT-5.2 produces a visually plausible silhouette but fails to preserve component fidelity, utilizing an incorrect piece inventory (e.g., hallucinating an extra square while omitting the parallelogram). We further observe that smaller models such as Qwen3-VL-8B-Instruct often arrange pieces in a simple linear sequence, indicating an inability to map symbolic coordinate specifications onto a 2D planar workspace.

5 Conclusion

We presented TangramPuzzle, a geometry-grounded benchmark specifically designed for evaluating MLLMs on compositional spatial reasoning. Addressing the limitations of existing semantic-focused evaluations, we introduced the Tangram Construction Expression (TCE), a symbolic geometric framework that enables the precise, machine-verifiable assessment of rigid body constraints and physical feasibility. The benchmark encompasses the Outline Prediction and End-to-End TCE Code Generation tasks, which jointly probe the spectrum of spatial intelligence. We conducted a comprehensive evaluation of open-source and closed-source MLLMs covering different scales, accompanied by in-depth exploratory analyses on in-context learning and modality dependency. By shifting the evaluation paradigm from approximate visual matching to exact constraint satisfaction, TangramPuzzle provides a standardized and challenging testbed for future research.

Limitations

While TangramPuzzle establishes a rigorous framework for evaluating compositional spatial reasoning through exact geometric constraints, several limitations remain. The benchmark is currently confined to 2D planar geometry with a fixed number of components, which may not fully capture the scalability and complexity of real-world 3D spatial manipulation tasks involving occlusion or variable object counts. Future work will seek to extend this evaluation paradigm to three-dimensional settings, thereby bringing the evaluation closer to real-world embodied scenarios.

Ethical Considerations

Potential Risks The TangramPuzzle benchmark focuses on abstract geometric reasoning and is constructed entirely from publicly available sources. Due to the inherently abstract nature of these geometric shapes, the dataset is devoid of personally identifiable information (PII), sensitive biometric data, or offensive content. Consequently, there are no foreseeable risks related to safety, discrimination, or surveillance.

Ethical Statement Our data construction process strictly adhered to ethical research guidelines. All source images were derived from public domains, ensuring no copyright infringement. During the human-in-the-loop annotation and validation stages, all participants were informed of the nature of the task and compensated fairly in accordance with local labor standards. Furthermore, as the dataset comprises solely rigid geometric shapes, it is inherently free from biometric data, social biases, or other privacy concerns. Thus, the proposed research direction and tasks are ethically benign and socially harmless. All data and annotations will be released under a permissive open-source license upon acceptance to facilitate transparency and reproducibility.

LLMs Usage Statement We use large language models and multimodal models as baselines to generate predictions for evaluation and analysis. During the writing of this paper, we also use large language models to assist with grammar checking and improving text fluency. The authors have carefully reviewed and verified all suggested revisions provided by the large language models to ensure their correctness, and take full responsibility for the content of this paper.

References

Anthropic. 2025. Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5. Accessed: 2026-01-04.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Laura Bofferding and Mahtob Aqazade. 2023. "where does the square go?": reinterpreting shapes when solving a tangram puzzle. *Educational Studies in Mathematics*, 112(1):25–47.

Laura Bofferding and Yi Zhu. 2023. Composing tangram puzzles to support shape transformation. *Mathematics Teacher: Learning and Teaching PK-12*, 116(7):503–510.

Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. 2023. Omni3d: A large benchmark and model for 3d object detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13154–13164.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024a. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

Song Chen, Xinyu Guo, Yadong Li, Tao Zhang, Mingan Lin, Dongdong Kuang, Youwei Zhang, Lingfeng Ming, Fengyu Zhang, Yuran Wang, and 1 others. 2025a. Ocean-ocr: Towards general ocr application via a vision-language model. *arXiv preprint arXiv:2501.15558*.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024b. Unified hallucination detection for multimodal large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3235–3252, Bangkok, Thailand. Association for Computational Linguistics.

Zixin Chen, Sicheng Song, Kashun Shum, Yanna Lin, Rui Sheng, Weiqi Wang, and Huamin Qu. 2025b. Unmasking deceptive visuals: Benchmarking multimodal large language models on misleading chart question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13767–13800.

Kanzhi Cheng, Wenpo Song, Jiaxin Fan, Zheng Ma, Qiushi Sun, Fangzhi Xu, Chenyang Yan, Nuo Chen, Jianbing Zhang, and Jiajun Chen. 2025a. CapArena:

Benchmarking and analyzing detailed image captioning in the LLM era. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14077–14094, Vienna, Austria. Association for Computational Linguistics.

Pengzhou Cheng, Zheng Wu, Zongru Wu, Tianjie Ju, Aston Zhang, Zhuosheng Zhang, and Gongshen Liu. 2025b. Os-kairos: Adaptive interaction for mllm-powered gui agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6701–6725.

Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, and 1 others. 2025c. Simplevqa: Multimodal factuality evaluation for multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4637–4646.

Ziming Cheng, Binrui Xu, Lisheng Gong, Zuhe Song, Tianshuo Zhou, Shiqi Zhong, Siyu Ren, Mingxiang Chen, Xiangchao Meng, Yuxin Zhang, and 1 others. 2025d. Evaluating mllms with multimodal multi-image reasoning benchmark. *arXiv preprint arXiv:2506.04280*.

Nianchen Deng, Lixin Gu, Shenglong Ye, Yinan He, Zhe Chen, Songze Li, Haomin Wang, Xingguang Wei, Tianshuo Yang, Min Dou, and 1 others. 2025. Internspatial: A comprehensive dataset for spatial reasoning in vision-language models. *arXiv preprint arXiv:2506.18385*.

Yihao Ding, Siwen Luo, Yue Dai, Yanbei Jiang, Zechuan Li, Geoffrey Martin, and Yifan Peng. 2025. A survey on mllm-based visually rich document understanding: Methods, challenges, and emerging trends. *arXiv preprint arXiv:2507.09861*.

Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. 2024. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*.

Wenlong Fang, Qiaofeng Wu, Jing Chen, and Yun Xue. 2025. guided mllm reasoning: Enhancing mllm with knowledge and visual notes for visual question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19597–19607.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, and 1 others. 2024a. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *arXiv preprint arXiv:2501.00321*.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024b. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

Google DeepMind. 2025. Gemini 3 pro. https://deepmind.google/models/gemini/pro/. Accessed: 2026-01-04.

Gavin Greif, Niclas Griesshaber, and Robin Greif. 2025. Multimodal llms for ocr, ocr post-correction, and named entity recognition in historical documents. *arXiv preprint arXiv:2504.00414*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Meng-Hao Guo, Jiajun Xu, Yi Zhang, Jiaxi Song, Haoyang Peng, Yi-Xuan Deng, Xinzhi Dong, Kiyohiro Nakayama, Zhengyang Geng, Chen Wang, and 1 others. 2025. Rbench: Graduate-level multidisciplinary benchmarks for llm & mllm complex reasoning evaluation. In *Forty-second International Conference on Machine Learning*.

Jiaxi Huang, Dongxu Wu, Hanwei Zhu, Lingyu Zhu, Jun Xing, Xu Wang, and Baoliang Chen. 2025a. Q-doc: Benchmarking document image quality assessment capabilities in multi-modal large language models. *arXiv preprint arXiv:2511.11410*.

Mingxin Huang, Yongxin Shi, Dezhi Peng, Songxuan Lai, Zecheng Xie, and Lianwen Jin. 2025b. Ocr-reasoning benchmark: Unveiling the true capabilities of mllms in complex text-rich image reasoning. *arXiv preprint arXiv:2505.17163*.

Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert Hawkins, and Yoav Artzi. 2022. Abstract visual reasoning with tangram shapes. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 582–601.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. 2025. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max W.F. Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research*, 2024.

Jiayi Kuang, Yinghui Li, Chen Wang, Haohao Luo, Ying Shen, and Wenhao Jiang. 2025. Express what you see: Can multimodal llms decode visual ciphers with intuitive semiosis comprehension? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12743–12774.

Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. The curse of

multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. *arXiv preprint arXiv:2410.12787*.

Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. 2025. Screenspot-pro: Gui grounding for professional high-resolution computer use. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 8778–8786.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Yinghui Li, Zishan Xu, Shaoshen Chen, Haojing Huang, Yangning Li, Shirong Ma, Yong Jiang, Zhongli Li, Qingyu Zhou, Hai-Tao Zheng, and 1 others. 2024. Towards real-world writing assistance: A chinese character checking benchmark with faked and misspelled characters. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8656–8668.

Jingping Liu, Ziyan Liu, Zhedong Cen, Yan Zhou, Yinan Zou, Weiyan Zhang, Haiyun Jiang, and Tong Ruan. 2025. Can multimodal large language models understand spatial relations? *arXiv preprint arXiv:2505.19015*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Fan Lu, Wei Wu, Kecheng Zheng, Shuailei Ma, Biao Gong, Jiawei Liu, Wei Zhai, Yang Cao, Yujun Shen, and Zheng-Jun Zha. 2025. Benchmarking large vision-language models via directed scene graph for comprehensive image captioning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19618–19627.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 2025a. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6924–6934.

Xinyu Ma, Ziyang Ding, Zhicong Luo, Chi Chen, Zonghao Guo, Derek F Wong, Xiaoyi Feng, and Maosong Sun. 2025b. Deepperception: Advancing r1-like cognitive visual perception in mllms for knowledge-intensive visual grounding. *arXiv preprint arXiv:2503.12797*.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2025. ChartQAPro: A more diverse and challenging benchmark for chart question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19123–19151, Vienna, Austria. Association for Computational Linguistics.

OpenAI. 2025. Introducing gpt-5.2. https://openai.com/index/introducing-gpt-5-2/. Accessed: 2026-01-04.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, and 1 others. 2025. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24838–24848.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. 2023. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3170–3180.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, and 1 others. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William E Bishop, Wei Li, Folawiyo Campbell-Ajala, and 1 others. 2025. Androidworld: A dynamic benchmarking environment for autonomous agents. In *The Thirteenth International Conference on Learning Representations*.

Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, and 1 others. 2025. Zerobench: An impossible visual benchmark for contemporary large multimodal models. *arXiv preprint arXiv:2502.09696*.

Fedor Rodionov, Abdelrahman Eldesokey, Michael Birsak, John Femiani, Bernard Ghanem, and Peter Wonka. 2025. Floorplanqa: A benchmark for spatial reasoning in llms using structured representations. *arXiv preprint arXiv:2507.07644*.

Kexian Tang, Junyao Gao, Yanhong Zeng, Haodong Duan, Yanan Sun, Zhening Xing, Wenran Liu, Kaifeng Lyu, and Kai Chen. 2025a. Lego-puzzles: How good are mllms at multi-step spatial reasoning? *arXiv preprint arXiv:2503.19990*.

Wei Tang, Yanpeng Sun, Qinying Gu, and Zechao Li. 2025b. Visual position prompt for mllm based visual grounding. *arXiv preprint arXiv:2503.15426*.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, and 1 others. 2025a.

Muirbench: A comprehensive benchmark for robust multi-image understanding. In *The Thirteenth International Conference on Learning Representations*.

Peijie Wang, Chao Yang, Zhong-Zhi Li, Fei Yin, Dekang Ran, Mi Tian, Zhilong Ji, Jinfeng Bai, and Cheng-Lin Liu. 2025b. Solidgeo: Measuring multimodal spatial math reasoning in solid geometry. *arXiv preprint arXiv:2505.21177*.

Wei Wang, Zhaowei Li, Qi Xu, Linfeng Li, YiQing Cai, Botian Jiang, Hang Song, Xingcan Hu, Pengyu Wang, and Li Xiao. 2025c. Advancing fine-grained visual understanding with multi-scale alignment in multi-modal models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14282–14301.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, and 1 others. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*.

Haoning Wu, Xiao Huang, Yaohui Chen, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025a. Spatialscore: Towards unified evaluation for multimodal spatial understanding. *arXiv preprint arXiv:2505.17012*.

Junda Wu, Yu Xia, Tong Yu, Xiang Chen, Sai Sree Harsha, Akash V Maharaj, Ruiyi Zhang, Victor Bursztyn, Sungchul Kim, Ryan A Rossi, and 1 others. 2025b. Doc-react: Multi-page heterogeneous document question-answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 67–78.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

xAI. 2024. Realworldqa. https://huggingface.co/datasets/xai-org/RealworldQA. Accessed: 2025-12-30.

Renqiu Xia, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Botian Shi, Junchi Yan, and Bo Zhang. 2025. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *IEEE Transactions on Image Processing*.

Han Xiao, Guozhi Wang, Yuxiang Chai, Zimu Lu, Weifeng Lin, Hao He, Lue Fan, Liuyang Bian, Rui Hu, Liang Liu, and 1 others. 2025. Ui-genie: A self-improving approach for iteratively boosting mllm-based mobile gui agents. *arXiv preprint arXiv:2505.21496*.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*.

Tianbao Xie, Jiaqi Deng, Xiaochuan Li, Junlin Yang, Haoyuan Wu, Jixuan Chen, Wenjing Hu, Xinyuan Wang, Yuhui Xu, Zekun Wang, and 1 others. 2025. Scaling computer-use grounding via user interface decomposition and synthesis. *arXiv preprint arXiv:2505.13227*.

Liangyu Xu, Yingxiu Zhao, Jingyun Wang, Yingyao Wang, Bu Pi, Chen Wang, Mingliang Zhang, Jihao Gu, Xiang Li, Xiaoyong Zhu, and 1 others. 2025a. Geosense: Evaluating identification and application of geometric principles in multimodal reasoning. *arXiv preprint arXiv:2504.12597*.

Peiran Xu, Sudong Wang, Yao Zhu, Jianing Li, and Yunjian Zhang. 2025b. Spatialbench: Benchmarking multimodal large language models for spatial cognition. *arXiv preprint arXiv:2511.21471*.

Rui Xu, Dakuan Lu, Zicheng Zhao, Xiaoyu Tan, Xintao Wang, Siyu Yuan, Jiangjie Chen, and Xu Yinghui. 2025c. Origamispace: Benchmarking multimodal llms in multi-step spatial reasoning with mathematical constraints. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yunqiu Xu, Linchao Zhu, and Yi Yang. 2025d. Mc-bench: A benchmark for multi-context visual grounding in the era of mllms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17675–17687.

Fernanda Miyuki Yamada, Harlen Costa Batagelo, João Paulo Gois, and Hiroki Takahashi. 2025. Tangan: solving tangram puzzles using generative adversarial network. *Applied Intelligence*, 55(6):1–27.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, and 1 others. 2025a. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*.

Songyuan Yang, Weijiang Yu, Wenjing Yang, Xinwang Liu, Huibin Tan, Long Lan, and Nong Xiao. 2025b. Wildvideo: Benchmarking lmms for understanding video-language interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jiakang Yuan, Tianshuo Peng, Yilei Jiang, Yiting Lu, Renrui Zhang, Kaituo Feng, Chaoyou Fu, Tao Chen, Lei Bai, Bo Zhang, and 1 others. 2025. Mme-reasoning: A comprehensive benchmark for logical reasoning in mllms. *arXiv preprint arXiv:2505.21327*.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, and 1 others. 2025. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15134–15186.

Z.ai. 2025. Glm-4.6v: Open source multimodal models with native tool use. https://z.ai/blog/glm-4.6v. Accessed: 2025-01-04.

Chao Zhang, Jiamin Tang, and Jing Xiao. 2024. Tangram: Benchmark for evaluating geometric element recognition in large multimodal models. *arXiv preprint arXiv:2408.13854*.

Xuanle Zhao, Xuexin Liu, Haoyue Yang, Xianzhen Luo, Fanhu Zeng, Jianling Li, Qi Shi, and Chi Chen. 2025. Chartedit: How far are mllms from automating chart analysis? evaluating mllms' capability via chart editing. *arXiv preprint arXiv:2505.11935*.

Yizhou Zhao, Liang Qiu, Pan Lu, Feng Shi, Tian Han, and Song-Chun Zhu. 2022. Learning from the tangram to solve mini visual tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 3490–3498.

Longtao Zheng, Zhiyuan Huang, Zhenghai Xue, Xinrun Wang, Bo An, and Shuicheng YAN. 2025a. Agentstudio: A toolkit for building general virtual agents. In *The Thirteenth International Conference on Learning Representations*.

Xu Zheng, Zihao Dongfang, Lutao Jiang, Boyuan Zheng, Yulong Guo, Zhenquan Zhang, Giuliano Albanese, Runyi Yang, Mengjiao Ma, Zixin Zhang, and 1 others. 2025b. Multimodal spatial reasoning in the large model era: A survey and benchmarks. *arXiv preprint arXiv:2510.25760*.

Liangyu Zhong, Fabio Rosenthal, Joachim Sicking, Fabian Hüger, Thorsten Bagdonat, Hanno Gottschalk, and Leo Schwinn. 2025. Focus: Internal mllm representations for efficient fine-grained visual question answering. *arXiv preprint arXiv:2506.21710*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Human Annotation Details

Recruitment and Payment To ensure the high quality and rigorous geometric precision of the TangramPuzzle benchmark, we invited researchers with backgrounds in geometry and computer science to conduct the data annotation and verification. We strictly adhered to local labor standards regarding compensation. The payment rate was determined based on the estimated time required to complete the annotation for each puzzle, ensuring that the hourly rate met or exceeded the local minimum wage requirements.

Annotation Process Annotators used our interactive tangram tool to assemble each target configuration by dragging, rotating, and flipping pieces. The annotation interface is shown in Figure 5. To avoid pixel-level misalignment, the tool incorporates a magnetic snapping mechanism that automatically aligns piece vertices to the global coordinate grid or to vertices of adjacent pieces, ensuring that spatial relationships are mathematically exact rather than visually approximate. After completing an instance, annotators exported the generated data, and we conducted additional visual inspection and re-annotation when necessary (e.g., minor misalignment, unintended overlaps, or disconnected components) to ensure geometric correctness.

Instructions and Consent Prior to the data collection process, all annotators were provided with comprehensive instructions detailing the usage of the annotation interface, the definition of the Tangram Construction Expression (TCE), and the strict physical constraints (e.g., non-overlap, boundary adherence). We obtained informed consent from all participants, who were explicitly informed that the annotated data would be used for academic research and released to the public.

B Detailed Evaluation Protocols

This section provides the formal definitions and precise calculation details for the metrics used in our evaluation.

B.1 Task 1 Metrics

Accuracy (Acc) Performance is measured using classification accuracy:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{a}_i = a_i)\,, \tag{1}$$

where $a_i \in \mathcal{A}$ denotes the ground-truth option for the i-th instance, $\hat{a}_i \in \mathcal{A}$ is the model's predicted option, $\mathcal{A}$ is the option set (e.g., $\{A, B, C, D\}$), and $\mathbb{I}(\cdot)$ is the indicator function.

Invalid Rate We further define the Invalid rate as the proportion of instances in which the model output does not yield any valid option that can be mapped to the candidate set. Invalid outputs may arise from formatting errors or from generating content unrelated to the candidate options. Formally:

$$\text{Invalid} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{a}_i \notin \mathcal{A})\,. \tag{2}$$

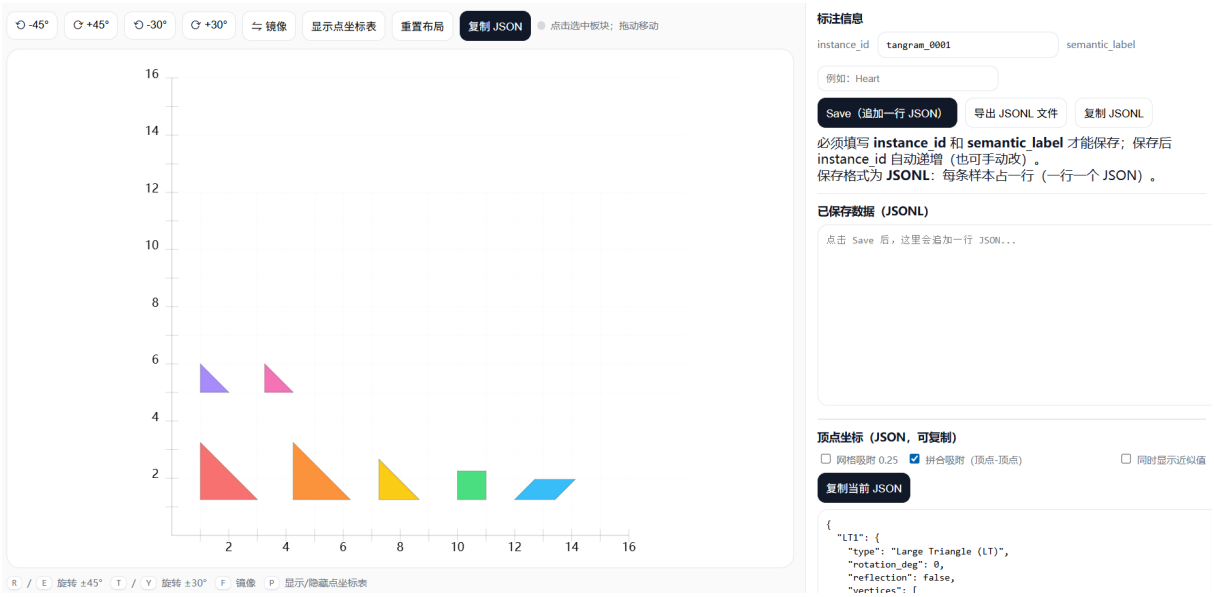

Figure 5: Annotation interface for TangramPuzzle.

B.2 Task 2 Metrics

Let $\mathcal{P} = \{P_1, \ldots, P_7\}$ denote the set of predicted piece polygons after assembly, and let $\mathcal{T}$ denote the ground-truth target silhouette. The evaluation pipeline consists of two stages: constraint validation and silhouette quality assessment.

Constraint Validation We consider three types of constraint errors:

- **TSE (Syntax Error)** checks whether the predicted output conforms to the required TCE format. A prediction is flagged as TSE if the JSON cannot be parsed, the number of pieces is incorrect, or piece types are invalid:

$$\mathrm{TSE} = \neg \mathrm{Valid}_{\mathrm{syn}}(\mathcal{P}), \tag{3}$$

  where $\mathrm{Valid}_{\mathrm{syn}}(\mathcal{P})$ denotes syntactic validity under the TCE schema.

- **RGE (Rigid Geometry Error)** verifies that each predicted piece preserves its original shape. For each piece, we compare its area and perimeter against the corresponding ground-truth specification. Let P_i be a simple polygon with ordered vertices $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. Its area is computed as:

$$\mathrm{Area}(P_i) = \frac{1}{2}\left|\sum_{k=1}^{n}(x_k y_{k+1} - x_{k+1} y_k)\right|, \tag{4}$$

  and its perimeter is given by

$$\mathrm{Perimeter}(P_i) = \sum_{k=1}^{n} \|\mathbf{v}_{k+1} - \mathbf{v}_k\|_2, \tag{5}$$

  where $\mathbf{v}_{n+1} = \mathbf{v}_1$. A rigid geometry error is triggered if either quantity deviates from that of the corresponding canonical tangram piece.

- **PE (Physical Error)** evaluates physical feasibility. A prediction is considered invalid if any two pieces overlap or if the union of all pieces is not a single connected shape.

Shape Similarity We adopt two complementary shape similarity metrics:

- **IoU (Intersection over Union)** measures the overlap between the predicted assembly U and the target silhouette $\mathcal{T}$:

$$\mathrm{IoU}(U, \mathcal{T}) = \frac{\mu(U \cap \mathcal{T})}{\mu(U \cup \mathcal{T})}, \tag{6}$$

  where $\mu(\cdot)$ denotes planar area.

- **Hausdorff Distance (d_H)** To capture boundary-level deviations, we compute the Hausdorff distance between the boundaries of U and $\mathcal{T}$:

$$d_H(\partial U, \partial \mathcal{T}) = \max\left\{ \sup_{x \in \partial U} \inf_{y \in \partial \mathcal{T}} \|x - y\|_2, \sup_{y \in \partial \mathcal{T}} \inf_{x \in \partial U} \|x - y\|_2 \right\}. \tag{7}$$

Here ∂U and $\partial \mathcal{T}$ denote the boundaries of the predicted assembly and the target silhouette, respectively, and $\|\cdot\|_2$ represents the standard Euclidean distance.

C Detailed Analytical Experiments

C.1 In-context Learning Details

We examine the effect of In-context Learning (ICL) by prepending three solved examples to the prompt for a subset of models (Qwen3-VL, GPT-5.2, Gemini3-Pro).

Syntactic Degradation. Contrary to the expectation that examples stabilize output, several models exhibited increased Syntax Error (TSE) rates compared to the zero-shot setting. For instance, Qwen3-VL-8B-Instruct and Gemini3-Pro showed higher failure rates in generating valid JSONs. This suggests that the extensive symbolic context imposes a cognitive load that interferes with the model's ability to strictly adhere to structural constraints.

Shape Refinement. Despite the syntactic instability, ICL yields clear benefits for silhouette-level quality in valid responses. Models demonstrated improved IoU and reduced Hausdorff distances, effectively learning to better approximate the global target shape. This suggests that while ICL serves as a potent mechanism for refining visual matching strategies and "teaching" the task format, it ultimately falls short of spontaneously instilling a fundamental understanding of strict physical validity in models that inherently lack it.

C.2 Textual Geometry Ablation

To quantify the models' dependency on textual metadata, we evaluated a Visual-Centric setting where explicit textual descriptions of the target outline (vertices and edges) were removed, leaving only the image with annotated coordinates.

Stability Collapse. The elimination of explicit textual outline specifications significantly destabilized structured generation. Notably, GPT-5.2 experienced a dramatic surge in syntax errors, with TSE rising from 7.19% to 39.67%. Similarly, the Deepseek series frequently hallucinated irrelevant content, indicating a heavy reliance on text prompts to guide the generation process.

Visual Grounding Gap. For the majority of models, the absence of textual data precipitated a marked deterioration in Silhouette Quality, exposing a critical dependency on semantic prompts. This confirms a widespread inability to precisely "read" geometric coordinates directly from raw visual inputs. However, Gemini3-Pro's robust performance in this setting serves as a notable exception, highlighting its advanced capability to perform rigorous reasoning solely based on visual grounding, independent of textual crutches.

D Detailed Human Performance Analysis

We evaluate human performance using three domain experts who independently solved the task sets under the same conditions as the models.

Performance Patterns As shown in our results, human solvers rarely violate fundamental geometric constraints such as piece shape preservation or piece count correctness. Consequently, human performance manifests as a binary outcome: a solution is either strictly valid or the participant fails to find a solution within a reasonable time. This stands in sharp contrast to model predictions, which often achieve partial visual similarity through impermissible physical violations (e.g., overlapping pieces or shape distortion).

Solving Strategies We observe substantial variance in success rates linked to the strategies employed. One expert achieved perfect performance by adopting a *systematic decomposition strategy*: all tangram pieces were interpreted as combinations of congruent isosceles right triangles (with unit legs and hypotenuse $\sqrt{2}$). By decomposing both the target silhouette and the pieces into this common geometric basis, the solver could explicitly reason about how composite shapes fit within the outline. In contrast, the other two experts primarily relied on *iterative trial-and-error*, repeatedly adjusting piece positions until a feasible configuration emerged. This heuristic approach resulted in lower success rates and significantly longer solving times.

Complexity Assessment Participants were asked to rate task difficulty on a scale from 1 to 5. The high average complexity score of 4.1 confirms that TangramPuzzle is non-trivial even for humans. On average, each instance required approximately 6.7 minutes to solve, reflecting the substantial cognitive effort required for precise geometric reasoning and assembly planning.