GamiBench: Evaluating Spatial Reasoning and 2D-to-3D Planning Capabilities of MLLMs with Origami Folding Tasks

Anonymous submission

Abstract

Multimodal large language models (MLLMs) are proficient in perception and instruction-following, but they still struggle with spatial reasoning: the ability to mentally track and manipulate objects across multiple views and over time. Spatial reasoning is a key component of human intelligence, but most existing benchmarks focus on static images or final outputs, failing to account for the sequential and viewpoint-dependent nature of this skill. To close this gap, we introduce *GamiBench*, a benchmark designed to evaluate spatial reasoning and 2D-to-3D planning in MLLMs through origami-inspired folding tasks. *GamiBench* includes 186 regular and 186 impossible 2D crease patterns paired with their corresponding 3D folded shapes, produced from six distinct viewpoints across three visual question-answering (VQA) tasks: predicting 3D fold configurations, distinguishing valid viewpoints, and detecting impossible patterns. Unlike previous benchmarks that assess only final predictions, *GamiBench* holistically evaluates the entire reasoning process of the models; measuring cross-view consistency, physical feasibility through impossible-fold detection and interpretation of intermediate folding steps. It further introduces new diagnostic metrics—viewpoint consistency (VC) and impossible fold selection rate (IFSR)—to measure how well models handle folds of varying complexity. By linking geometric evaluation with sequential reasoning, *GamiBench* enables a comprehensive evaluation of state-of-the-art MLLMs, revealing significant limitations in spatial reasoning capabilities and creating a new pipeline to advance geometric understanding in real-world contexts. The *GamiBench* dataset and code will be made available upon publication.

1 Introduction

Spatial reasoning is a fundamental component of human intelligence, crucial for interacting with the physical world, understanding relationships between objects, and executing multi-step actions. Tasks such as building furniture or folding origami require mentally simulating spatial transformations and tracking changing object states. As the interactions between artificial intelligence systems and real-world environments evolve, developing models that can reason about space and change has become a core challenge. Recent advances in multimodal large language models (MLLMs) show strong progress in image recognition, VQA, and instruction following (Dongfang et al. 2025), (Jiang et al. 2025); however, they clearly struggle on high-quality and temporally extended sequential spatial reasoning tasks (Rajabi and Kosecka 2024), (Tang et al. 2025), (Valmeekam et al. 2023).

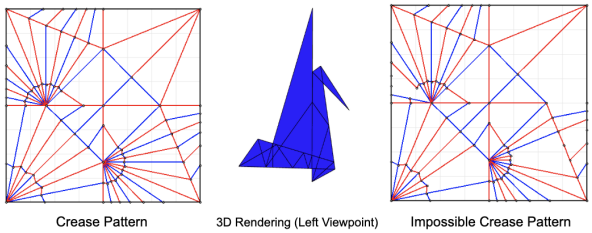

Figure 1: 2D normal crease pattern (left), 3D final fold state (middle), and 2D impossible crease pattern (right) of an example Crane variation (*furutaoripa_orihazuru*).

Existing multimodal benchmarks and tools have advanced evaluation breadth (e.g., perception, captioning, instruction following) (Fu et al. 2023), (Liu et al. 2023), but typically emphasize single-step judgments, static scenes, or end-state accuracy (Fu et al. 2023). Specialized efforts in spatial or procedural reasoning test important skills such as geometric consistency (Tang et al. 2025), (Rajabi and Kosecka 2024), arrangement from descriptions (Tang et al. 2025), or plan generation (Valmeekam et al. 2023), but often isolate narrow capabilities, fix visual formats, or under-specify multi-view consistency and temporal coherence. Thus, currently there is no framework that holistically evaluates how models plan, update, and validate spatial states across time and across multiple viewpoints of the same 3D structure.

In this work, we introduce *GamiBench*, a novel benchmark and framework for sequential spatial planning and reasoning that uses origami-inspired tasks to map 2D crease patterns to 3D multi-view final states (front, back, top, bottom, left, right). We take inspiration from the common traditional Japanese art of paper folding, Origami, using a single square sheet of paper to create a figure without cutting, gluing, or marking. The assembly of the 3D Origami model consists of dozens of intermediary discrete folds to transform a 2D paper plane into a 3D structure, providing an ideal foundation for testing sequential spatial reasoning abilities. We carefully curate a dataset of 186 regular and 186 impossible 2D square crease pattern configurations and 372

adjacent 3D folds, where each fold is part of a pair of correct folds for a given 2D crease pattern, using existing online tools like Oriedita (Oolbekkink et al. 2021) and Origami Simulator (Ghassaei 2018). *GamiBench* encompasses a collection of 744 Visual Question-Answering (VQA) multiple-choice sets spanning across 3 tasks, one of which is conditioned on the final answer, from this dataset.

First, we develop a set of tests to assess MLLMs' basic spatial transformation understanding capabilities, including single-step 2D-to-3D mapping, reinforcement of initial correctness via alternative 3D final states, and assessment of impossible fold configuration acknowledgement. We develop our own metrics, such as Viewpoint Consistency (VC) and Impossible Fold Selection Rate (IFSR), to measure model performance across these 3 tasks.

GamiBench provides several distinct advantages compared to existing spatial understanding benchmarks by evaluating (1) multi-perspective consistency across 3D views, (2) feasibility via impossible-fold detection (violations of physical origami axioms), and (3) sequential interpretation of intermediate states induced by textual instructions and visual transitions. Furthermore, we classify folds by complexity to stress-test models' robustness to geometric density.

Leveraging *GamiBench*, we conduct evaluations on 21 state-of-the-art MLLMs, including proprietary models such as GPT-5 and Gemini-2.5-Flash, as well as leading open-source alternatives such as Llama-4-Maverick, Gemma-3-27B-IT, Cogito-V2-Preview-Llama-109B-MoE, and GLM-4.5V. Our results reveal a substantial gap in MLLM spatial reasoning proficiency. Even the strongest models struggle with basic spatial understanding tasks, at times performing worse than weaker models on the same task.

Concretely, our contributions are as follows:

- A multi-view, sequential spatial benchmark that evaluates 2D-to-3D reasoning beyond final-state accuracy, with tasks that require coherence across six views and across time.
- New evaluation axes—Viewpoint Consistency and Impossible Fold Selection Rate (IFSR)—that diagnose failure modes missed by standard accuracy metrics.
- Origami-inspired task suite combining textual instructions with visual states to test whether MLLMs can plan, update, and verify geometric transformations.
- Complexity controls (simple vs. complex crease patterns) enabling analysis of scale effects on spatial planning.

Together, these contributions to spatial evaluation from static recognition toward procedural, multi-view reasoning provide clearer signals about where current MLLMs fall short and how future architectures or training regimes might improve.

2 Related Work

Various comprehensive benchmarks have been introduced to assess various multimodal capabilities.

Foundations of Spatial and Embodied Reasoning. Spatial intelligence has long been identified as a fundamental component of human cognition, underlying reasoning about geometry and physical relationships (Bornstein 1986). Recent embodied MLLM frameworks extend this concept to machine perception and control. OpenVLA (Kim et al. 2024) links vision, language, and action to achieve generalist visuomotor manipulation across diverse robots, while ManipLLM (Li et al. 2024) integrates affordance reasoning and pose prediction for object-centric manipulation in real environments. In autonomous driving, DriveMLLM (Guo et al. 2024) benchmarks spatial scene understanding under occlusion and dynamic layouts, and DriveMLM (Wang et al. 2023) aligns multimodal perception with behavioral-planning states for closed-loop navigation. Follow-up studies reveal that modality imbalance limits generalization from simple to complex visual reasoning tasks (Park et al. 2025), and that MLLMs still struggle to remember and reconstruct 3D spaces from sequential observations (Yang et al. 2024). Collectively, these works highlight the gap between embodied perception and sustained geometric reasoning across changing viewpoints.

General Multimodal Evaluation Benchmarks. Comprehensive multimodal benchmarks have advanced large-scale evaluation of perception and reasoning. MME (Fu et al. 2023) systematically measures 14 multimodal subtasks, exposing persistent weaknesses such as object hallucination and spatial reasoning errors. MMBench (Liu et al. 2023) introduces a bilingual multiple-choice framework for fine-grained multimodal assessment, while MMMU (Yue et al. 2024) extends difficulty to college-level, discipline-specific visual reasoning. Despite their breadth, these datasets primarily test static understanding and lack mechanisms for evaluating multi-view geometric coherence or physical feasibility, both of which are core aspects addressed by *GamiBench*.

Spatial and Geometric Reasoning. Recent benchmarks directly target spatial reasoning capabilities. GSR-Bench (Rajabi and Kosecka 2024) evaluates object-relation understanding and shows that MLLMs frequently confuse depth and relative position, revealing weak geometric grounding. LEGO-Puzzles (Tang et al. 2025) probes multi-step spatial reasoning through LEGO-based assembly tasks and uncovers severe performance gaps between humans ($\sim$90%) and MLLMs ($\sim$50%). 3DSRBench (Ma et al. 2024) assesses 3D reasoning across orientation, occlusion, and viewpoint changes, finding that accuracy drops sharply under non-canonical perspectives. OSR-Bench (Dongfang et al. 2025) examines omnidirectional spatial reasoning over 360° panoramic inputs and reports poor rotation invariance in current models. Psychometric analysis of basic spatial abilities (Xu et al. 2025) corroborates these findings, identifying deficits in mental rotation and coordinate transformation. Zhang et al. (Zhang et al. 2025) argue that such weaknesses cannot be solved through scaling alone and call for geometry-aware, physically grounded training data. *GamiBench* complements these efforts by uniting physical validity and cross-view spatial consistency through origami-inspired folds.

Sequential Planning and Temporal Reasoning. Beyond static geometry, MARBLE (Jiang et al. 2025) tests multimodal reasoning and planning across multi-step spa-

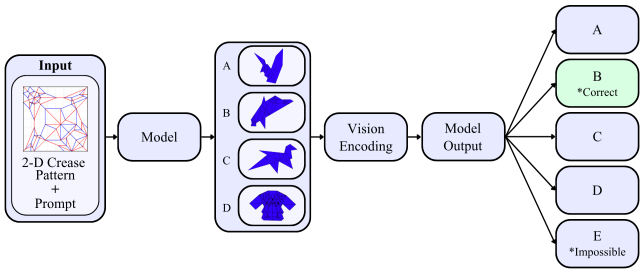

Figure 2: Visual mapping of the task flow in GamiBench. The model receives a 2D crease pattern (*komatsu_dolphin*) and text prompt, encodes candidate 3D folds, and outputs the most plausible match among multiple choices.

tial tasks, showing near-random performance even in simplified subtasks. PlanBench (Valmeekam et al. 2023) provides a textual framework for evaluating reasoning about actions and change, serving as a foundation for later multimodal planning benchmarks. Similarly, *GamiBench* models sequential folding as a structured planning problem, requiring consistency across transformations and temporal stages.

Interpretability and Evaluation Bias. VERIFY (Bi et al. 2025) isolates visual reasoning fidelity using human-annotated reasoning paths, revealing that models often reach correct answers for the wrong visual evidence. In parallel, Zheng et al. (Zheng et al. 2024) expose a systemic multiple-choice selection bias in LLMs, showing preference for specific option positions independent of content. *GamiBench* incorporates an Impossible Fold Selection Rate (IFSR) metric and balanced MCQ design to mitigate such biases, ensuring that geometric reasoning rather than positional heuristics drives model performance.

Collectively, these studies have broadened multimodal reasoning evaluation yet remain restricted to static perception, symbolic planning, or 2D scene understanding. None directly assess 2D-to-3D transformation, multi-view spatial coherence, or physical feasibility. *GamiBench* addresses this gap through origami-inspired tasks that couple 2D crease patterns with multi-view 3D structures, employing Viewpoint Consistency (VC) and Impossible Fold Selection Rate (IFSR) to capture how MLLMs reason about geometry, transformation, and physical constraints.

3 Methodology

GamiBench is a multimodal benchmark that evaluates MLLMs on two-dimensional-to-three-dimensional spatial reasoning through origami-inspired tasks. Designed to evaluate both perception and procedural planning, *GamiBench* formalizes spatial reasoning as the process of mapping crease patterns to structures while maintaining geometric consistency throughout multiple viewpoints.

3.1 Task Definition

GamiBench makes use of two task clusters showing differing reasoning regimes: (i) Single-Step Spatial Understanding (SSSU) that measures an MLLM's ability to infer a folded 3D structure based on a single 2D crease pattern, and (ii) Multi-Step Spatial Reasoning that requires temporal reasoning within the context of cross-view coherence.

For the one-step setup, models will produce final shape identification and view point recognition. The multi-step setting will extend this reasoning over to complex folding progressions, requiring models to be cognizant of geometric transitions and determine the step at which a fold can or cannot feasibly take place; thus exposing the models understanding of spatial continuity and physical possibility. This multi-step setting is *implicit*; we do not prompt models with intermediate folding steps. Our multi-step process tests models' multi-step understanding of complex folds with 40+ creases that require lots of spatial transformations. An overview of the task formulation and model decision flow is shown in Figure 2. For analysis of scaling behavior, folds are classified into two levels; simple folds (less than 40 creases), and complex folds (greater than or equal to 40 creases).

3.2 Dataset Curation

Dataset Collection. The dataset comprises 186 origami instances, each containing a normalized 2D crease pattern, 2D impossible crease pattern, and two 3D renderings captured from canonical viewpoints. For each instance, only two of the six canonical viewpoints—front, back, top, bottom, left, and right—are selected and verified by humans for plausibility. To construct these instances, we sourced crease patterns from the Flat-Folder platform (Ku 2025), which provides publicly available flat-foldable crease configurations, and verified their geometry using the open-source software Oriedita (Oolbekkink et al. 2021), which enables precise control over mountain–valley assignments and fold geometry. Each pattern was then simulated and exported as a 3D folded mesh via Origami Simulator (Ghassaei 2018). All meshes were rendered under identical lighting and camera conditions to ensure viewpoint consistency. We include both physically valid and constraint-violating (impossible) folds to evaluate model sensitivity to geometric feasibility. Each instance was normalized and aligned to a shared coordinate frame to maintain comparability across samples.

Foldability Verification. We verify the feasibility of each crease pattern using Oriedita's built-in programmatic verifier, CAMv (Oolbekkink et al. 2021), which automatically applies flat-foldability constraints derived from origami axioms and theorems such as Kawasaki's (Barile 2002) and Maekawa's (Maekawa 1983). This ensures that physically valid folds satisfy all geometric consistency conditions, while infeasible folds violate at least one constraint. The verifier's outputs are cross-checked during data generation to confirm that impossible folds arise from genuine geometric contradictions rather than rendering artifacts.

Origami Axioms. To distinguish physically valid folds from infeasible ones, we define fold legality according to geometric and physical origami constraints. A fold is considered feasible if it satisfies flat-foldability conditions such as Kawasaki's (Barile 2002) and Maekawa's theorems (Maekawa 1983) and maintains continuous paper geometry without self-intersection. Impossible folds are generated by deliberately violating these principles, for example, by enforcing inconsistent mountain–valley assignments

Task-Specific Template (here for task *SSSU*)

Instruction: You are an **Origami Folding Expert**. You will be given the final crease pattern of a folded origami model and four candidate 3D models labeled A through D. Evaluate **all four** symmetrically—do not privilege any order. Only one of the candidate models corresponds exactly to the result of folding the given crease pattern. In the crease pattern, red lines represent *mountain* folds and blue lines represent *valley* folds. Your task is to analyze the crease pattern and select the correct 3D model based solely on visual and geometric reasoning. Consider fold types, symmetry, flap orientation, and structural features visible in the crease pattern. If none of the four models are possible, respond with option **E** ("This fold is impossible"). At each stage, respond with a single uppercase letter **A**, **B**, **C**, **D**, or **E**.

Question: What is the correct 3D model for the given 2D crease pattern?

Answer: Ground Truth

Figure 3: Task-specific template. Our QA template includes instructions, question, and answer for the *SSSU* task.

or overlapping crease intersections that would cause self-intersections in a real sheet. Through this structured design, *GamiBench* will test MLLMs understanding of spatial transformations and ability to maintain consistency across multi-view geometric representations. For completeness, we include the Huzita–Hatori axioms (Huzita 1991; Hatori 2002), which define the geometric foundations of origami foldability. These axioms state that a fold can pass through any two given points, or place one point onto another; that a fold can align two lines or pass through a point while being perpendicular to a given line; and that more complex conditions allow folds to place one point onto a line while passing through another point, to map two points each onto their respective lines, or to position a point onto one line while remaining perpendicular to another. Collectively, these seven axioms describe the complete set of single-fold operations possible under Euclidean geometry and serve as the basis for distinguishing valid from infeasible fold structures in our dataset.

Question-Answer Generation and Quality Control. To streamline our evaluations of MLLMs, we design a template for question-answer generation. See Figure 3 for an example. Each data example includes a multimodal instruction, multiple-choice prompt, and an answer. For each 2D data example, we define its 3D fold label as the correct answer, randomly assigned to a letter between A-D. Additionally, we randomly sample 3 other 3D viewpoints in our dataset as incorrect answers, assigned to the remaining letters to build our multiple-choice answer bank. Our prompt remains immutable throughout all 3 tasks (See Figure 3). We condition the viewpoint consistency task on the event that the standard task is correctly answered.

To maintain reproducibility and minimize duplication errors, five human annotators reviewed each multiple-choice set, verifying (1) the absence of duplicated or mirrored folds, (2) the balanced difficulty across options, and (3) the lack of visually ambiguous distractors. These checks ensured that the final answer banks remained balanced, non-redundant, and free from biases that could influence model responses. We also set a seed of 42 for randomizing answer choices and a temperature of 0 for most models to facilitate deterministic outputs.

4 Results

4.1 Overall Performance

Among the 21 MLLMs evaluated in *GamiBench*, performance varied significantly, particularly when transitioning from simple to complex folding tasks (See Table 1). Most models demonstrated reasonable competence in SSSU, such as interpreting a 2D crease pattern and identifying its corresponding 3D structure. However, accuracy declined sharply when multi-step reasoning was required. Tasks that demanded tracking fold sequences or maintaining geometric consistency across temporal stages were especially challenging, revealing limitations in the model's ability to integrate spatial transformations over time.

Evaluation Metrics. To ensure consistent assessment across all three VQA tasks, *GamiBench* employs three complementary measurements: Accuracy, Impossible Fold Selection Rate (IFSR), and Viewpoint Consistency (VC). Accuracy captures the model's ability to identify the correct 3D fold or viewpoint among four multiple choice options, providing a direct measure of categorical prediction quality. IFSR quantifies how often a model incorrectly labels a valid fold as impossible, reflecting its sensitivity to geometric infeasibility and physical constraint reasoning. VC measures whether a model that correctly identifies the 3D fold for a crease pattern from one viewpoint remains correct when the same 3D fold is presented from a different viewpoint (randomly chosen from the remaining available views), using the same 3 distractor candidates as in the primary trial. Thus, VC here is a *conditional* single-view re-test accuracy: the fraction of eligible items whose correctness persists under a viewpoint change with an unchanged candidate set except for the new correct image (the denominator is the number of primary successes). High VC scores indicate consistent multi-view understanding, while lower scores reveal discrepancies in spatial alignment or rotation tracking.

Let $\mathcal{D}$ be the set of normal crease patterns with a valid 3D fold. For item $i \in \mathcal{D}$, let y_i be the correct 3D fold, v_0 the primary view, and $v_1 \neq v_0$ the re-test view (same three distractors).

Normal accuracy.

$$\mathrm{Acc} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \mathbf{1}\left[\hat{y}_i^{(v_0)} = y_i \right] \tag{1}$$

Define the subset of primary-view successes

$$\mathcal{S} = \left\{ i \in \mathcal{D} : \hat{y}_i^{(v_0)} = y_i \right\}. \tag{2}$$

Conditional viewpoint consistency.

$$\mathrm{VC} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathbf{1}\left[\hat{y}_i^{(v_1)} = y_i \right] \tag{3}$$

Model	Simple				Complex				Overall	
	Normal	VC	IFSR	Imp.	Normal	VC	IFSR	Imp.	Normal	Imp.
Closed Source										
Claude Opus 4.1	44.9	84.6	22.2	14.3	38.7	84.2	22.4	11.4	41.8	12.9
Claude Opus 4	38.8	63.2	34.1	12.2	36.5	64.1	34.9	15.9	37.7	14.1
Claude 4.5 Sonnet	34.7	82.0	20.6	0.0	48.2	88.2	19.0	16.7	41.5	8.4
Grok-4-Fast	36.7	70.7	29.4	0.0	33.6	64.0	31.7	6.1	35.2	3.1
GPT-o3	44.9	80.7	26.2	8.2	38.0	81.9	25.4	9.8	41.5	9.0
GPT-5	67.3	88.4	25.4	6.1	60.6	88.9	24.6	10.6	64.0	8.4
GPT-4o	61.2	83.9	27.0	8.2	38.7	82.2	28.6	18.2	50.0	13.2
GPT-4o-Mini	26.5	84.6	22.2	22.4	34.3	85.7	21.4	34.1	30.4	28.3
Gemini-2.0-Flash	26.5	83.5	22.2	10.2	29.9	85.0	23.0	25.8	28.2	18.0
Gemini-2.5-Pro	49.0	90.3	15.9	8.2	43.8	91.2	15.1	12.1	46.4	10.2
Gemini-2.5-Flash	49.0	88.3	17.5	8.2	38.0	89.5	16.7	5.3	43.5	6.8
Open Source										
Mistral-Medium-3.1	26.5	84.6	22.2	42.0	24.8	82.5	24.6	40.2	25.7	41.1
Llama-4–Scout	34.7	64.1	26.2	0.0	44.5	67.4	26.2	0.0	39.6	0.0
Llama-4–Maverick	24.5	59.2	23.0	0.0	43.8	59.0	29.4	0.8	34.2	0.4
Gemma-3–27B-IT	38.8	86.2	23.0	12.0	46.0	86.0	22.2	8.3	42.4	10.2
Cogito-V2-Preview Llama-109B-MoE	49.0	88.1	22.2	6.0	39.4	78.3	22.2	15.9	44.2	11.0
Qwen3-VL-8B–Thinking	28.6	67.6	22.2	0.0	27.0	64.0	23.0	0.0	27.8	0.0
Qwen3-VL-30B-A3B–Thinking	26.5	66.1	24.6	0.0	32.8	69.0	24.6	0.8	29.7	0.4
Qwen3-VL-235B-A22B–Thinking	24.5	66.7	22.2	2.0	35.8	72.9	22.2	0.8	30.2	1.4
Microsoft Phi-4 Multimodal-Instruct	36.7	66.7	27.0	2.0	38.7	72.9	23.8	3.0	37.7	2.5
GLM-4.5V	46.9	78.3	23.0	6.0	40.1	79.4	23.8	4.5	43.5	5.3

Table 1: GamiBench results (percent). Two-level headers: Simple vs. Complex; rightmost columns macro-average across both complexities. VC = viewpoint consistency. Imp. = Impossible. Overall best, second best, and third best are highlighted as such.

4.2 Closed-Source Models

Among closed-source systems, GPT-5 stood out as the strongest performer. It correctly identified 60.6% of complex folds and 67.3% of simple folds, while maintaining relatively high viewpoint consistency and angular stability (VC: 65.1% and 69.7%). These results suggest that GPT-5 has a comparatively stable understanding of 3D spatial relationships across multiple views.

In contrast, Grok-4-Fast and Gemini-2.0-Flash were the least effective in this group, with complex-fold accuracies of 33.6% and 29.9%, respectively. Claude Opus 4.1 and Claude 4.5 Sonnet achieved midrange results, with complex-fold accuracies between 38.7% and 48.2%, though the latter model showed stronger spatial coherence (VC up to 82.4%). GPT-4o performed well on simpler folds (61.2%) but plateaued on complex ones (38.7%), a trend consistent across most closed-source models.

Across this group, the Impossible Fold Selection Rate (IFSR), which measures how often a model incorrectly classifies a valid fold as impossible, remained consistently low, typically between 0–14%. Although a low rate might initially appear favorable, it actually reveals a limited sensitivity to geometric infeasibility: models rarely identify impossible folds even when they should. This suggests that despite strong visual reasoning, most closed-source MLLMs still lack a robust understanding of physical constraints and spatial validity.

4.3 Open-Source Models

Open-source models exhibited wider variability but often performed surprisingly well relative to their commercial counterparts. Llama-4-Scout, Llama-4-Maverick, and Gemma-3-27B-IT reached complex-fold accuracies between 44–46%, matching or slightly exceeding several closed-source systems. More impressively, their VC scores frequently surpassed 80%, suggesting a strong ability to maintain visual and geometric consistency across viewpoints, even when their final predictions were not always correct.

Cogito-V2-Preview-Llama-109B-MoE balanced accuracy and coherence particularly well, achieving 39.4% on complex folds and 49.0% on simple folds, with outstanding VC performance (85–92%). On the other hand, the Qwen3-VL models (8B, 30B, and 235B) consistently underperformed, recording accuracies below 33% on complex folds.

The GLM-4.5V model emerged as one of the most ca-

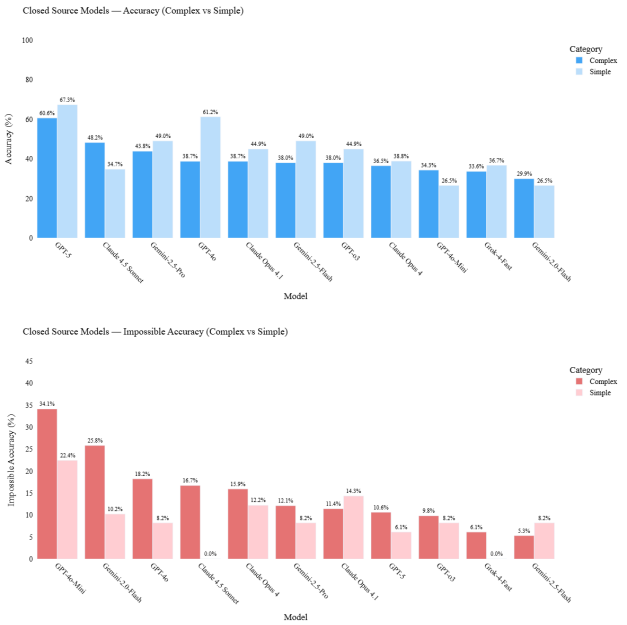

Figure 4: GamiBench Closed-Source Evaluations. Regular accuracy (top) and Impossible accuracy (bottom) of models, sorted in descending order from left to right, complex and simple.

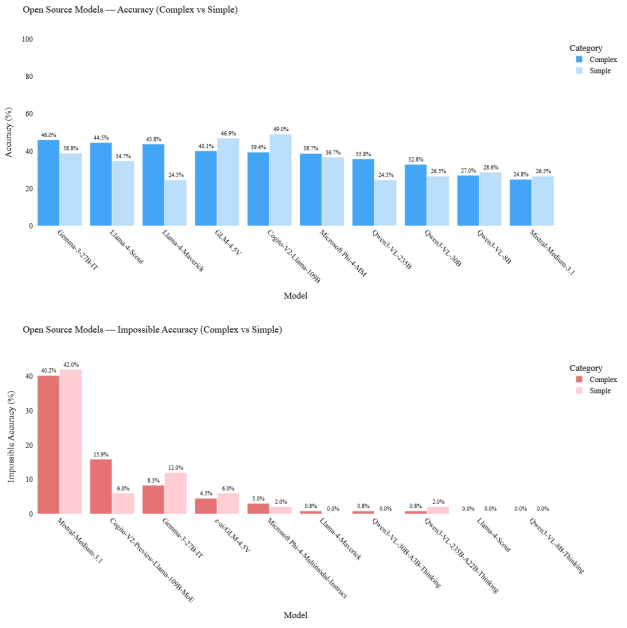

Figure 5: GamiBench Open-Source Evaluations. Regular accuracy (top) and Impossible accuracy (bottom) of models, sorted in descending order from left to right, complex and simple.

pable open systems, achieving 40.1% accuracy on complex folds and 46.9% on simple ones. Its VC scores (87.3% and 78.3%) placed it among the most consistent models in cross-view reasoning, showing that open models can indeed compete on structural coherence even without proprietary optimization.

4.4 Trends and Observations

Across all systems, task complexity was the clearest predictor of performance. On average, models performed 10–15% better on simple folds than on complex ones. Complexity controls in *GamiBench* categorize folds according to geometric density and reasoning depth. For our purposes, simple folds (fewer than 40 creases) primarily test localized spatial mapping and single-step 2D-to-3D transformations. In contrast, complex folds (40 or more creases) require multi-step reasoning, constraint tracking, and cross-view geometric coherence. Interpreting our results through this framework shows that increasing geometric density substantially amplifies cognitive and planning demands. While most MLLMs maintain stable accuracy on simple folds, their performance degrades as structural interactions grow combinatorially. This trend validates *GamiBench*'s complexity controls as a diagnostic tool for assessing how model reasoning scales with geometric and procedural difficulty. However, higher viewpoint consistency did not always imply correctness. Some models produced geometrically consistent yet incorrect shapes, revealing what we term a visual plausibility bias. In such cases, the models could maintain coherent visuals while misunderstanding the underlying 3D geometry.

Another persistent limitation was the inability to distinguish between valid and impossible folds. Even top-performing models like Mistral-Medium-3.1 and GPT-4o-Mini achieved combined average IFSRs of 27.6%, and 14.8%, respectively. This confirms that reasoning about spatial constraints and physical feasibility remains an open challenge in multimodal modeling.

We observe a miscalibration in Llama-4–Scout and Llama-4–Maverick: both select "impossible" 0% of the time on truly impossible items, yet show IFSR = 26.2%, indicating they sometimes (incorrectly) choose "impossible" on normal items. This asymmetric error pattern reflects a decision bias against the "E" option under ground-truth impossible conditions, limiting reliability on feasibility detection.

Finally, we observed a subtle but consistent selection bias in multiple-choice evaluations. Several models showed a tendency to favor specific answer options (e.g., "Option A") regardless of content, echoing patterns noted in prior work (Zheng et al. 2024). Future benchmark iterations should further mitigate this issue by incorporating generative, open-response formats that reduce residual positional bias and more effectively capture authentic reasoning ability.

4.5 Limitations

While *GamiBench* provides a structured and interpretable framework for assessing 2D-to-3D spatial reasoning, several limitations remain. The benchmark focuses on synthetic

origami-inspired folds and does not yet account for real-world conditions such as material deformation, lighting variation, or visual clutter, which may limit its generalization beyond crease-pattern reasoning. We also do not conduct *true* multi-step tasks such as predicting intermediary folds between 2D and 3D. The dataset size is moderate, and its complexity definition, which relies primarily on crease count, may not accurately reflect true planning difficulty caused by symmetry or long-range geometric dependencies. This is further supported by our finding that some models performed better on complex folds than on simple ones (i.e. Llama-4 series, Claude 4.5 Sonnet), indicating that crease count alone does not fully capture the true planning difficulty.

Furthermore, evaluations were performed under a single prompt and decoding configuration without systematic temperature tuning or standardized API normalization. Because proprietary APIs change over time, reproducibility remains partially limited. Future work could establish a human baseline to contextualize model performance, increase data realism and scale, introduce interactive folding and feedback-based tasks, integrate calibrated uncertainty metrics, and standardize evaluation protocols to improve comparability across evolving MLLM architectures.

5 Conclusion

Thus, we introduce *GamiBench*, a new benchmark for evaluating spatial reasoning and 2D-to-3D planning that aims to push the limits of modern MLLMs. By coupling origami-inspired crease patterns with multi-view 3D states, *GamiBench* advances beyond static visual understanding toward dynamic and multi-perspective spatial reasoning. Our results show that although leading MLLMs exhibit emerging competence in single-step geometric inference, they struggle to maintain spatial coherence across time, viewpoints, and physical constraints. Looking ahead, we envision *GamiBench* as a foundation for future spatial reasoning research, encouraging the development of models that can truly think in space by integrating perception, geometry, and physical reasoning into a unified understanding of real-world dynamics.

References

Barile, M. 2002. Kawasaki's Theorem. MathWorld—A Wolfram Resource.

Bi; et al. 2025. VERIFY: A Benchmark of Visual Explanation and Reasoning for Investigating Multimodal Reasoning Fidelity. *arXiv preprint arXiv:2503.11557*.

Bornstein, M. H. 1986. The Psychology of Spatial Intelligence: Reasoning about Geometry and Physical Relationships. *Journal of Cognitive Development*.

Dongfang, Z.; et al. 2025. Are Multimodal Large Language Models Ready for Omnidirectional Spatial Reasoning? *arXiv preprint arXiv:2505.11907*.

Fu; et al. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*.

Ghassaei, A. 2018. Origami Simulator [Computer software].

Guo, X.; et al. 2024. DriveMLLM: A Benchmark for Spatial Understanding with Multimodal Large Language Models in Autonomous Driving. *arXiv preprint arXiv:2409.01587*.

Hatori, K. 2002. Introduction to the Huzita–Hatori Axioms. Proceedings of the 3rd International Meeting of Origami Science, Mathematics, and Education (3OSME), Montréal, Canada.

Huzita, H. 1991. The Axioms of Origami. Proceedings of the First International Meeting of Origami Science and Technology, Ferrara, Italy.

Jiang, Y.; et al. 2025. MARBLE: A Hard Benchmark for Multimodal Spatial Reasoning and Planning. *arXiv preprint arXiv:2506.22992*.

Kim, M. J.; et al. 2024. OpenVLA: An Open-Source Vision-Language-Action Model. *arXiv preprint arXiv:2406.09246*.

Ku, J. S. 2025. Flat-Folder: Software to Compute and Analyze Flat-Foldable States of Crease Patterns. GitHub repository.

Li, X.; et al. 2024. ManipLLM: Embodied Multimodal Large Language Model for Object-Centric Robotic Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18061–18070.

Liu; et al. 2023. MMBench: Is Your Multimodal Model an All-Rounder? *arXiv preprint arXiv:2307.06281*.

Ma, W.; et al. 2024. 3DSRBench: A Comprehensive 3D Spatial Reasoning Benchmark. *arXiv preprint arXiv:2412.15492*.

Maekawa, J. 1983. Maekawa's Theorem. MathWorld—A Wolfram Resource.

Oolbekkink, G.; et al. 2021. Oriedita [Computer software]. GitHub repository.

Park, S.; et al. 2025. Generalizing from Simple to Hard Visual Reasoning: Mitigating Modality Imbalance in Vision-Language Models. *arXiv preprint arXiv:2501.02669*.

Rajabi, N.; and Kosecka, J. 2024. GSR-Bench: A Benchmark for Grounded Spatial Reasoning Evaluation via Multimodal LLMs. *arXiv preprint arXiv:2406.13246*.

Tang, K.; et al. 2025. LEGO-Puzzles: How Good Are MLLMs at Multi-Step Spatial Reasoning? *arXiv preprint arXiv:2503.19990*.

Valmeekam, K.; et al. 2023. Planbench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning About Change. In *Advances in Neural Information Processing Systems*, volume 36, 38975–38987.

Wang, W.; et al. 2023. DriveMLM: Aligning Multi-Modal Large Language Models with Behavioral Planning States for Autonomous Driving. *arXiv preprint arXiv:2312.09245*.

Xu, W.; et al. 2025. Defining and Evaluating Visual Language Models' Basic Spatial Abilities: A Perspective from Psychometrics. *arXiv preprint arXiv:2502.11859*.

Yang, J.; et al. 2024. Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces. *arXiv preprint arXiv:2412.14171*.

Yue, X.; et al. 2024. MMMU: A Massive Multimodal Multidisciplinary Understanding Benchmark. *arXiv preprint arXiv:2406.08669*.

Zhang; et al. 2025. A Call for New Recipes to Enhance Spatial Reasoning in MLLMs. *arXiv preprint arXiv:2504.15037*.

Zheng; et al. 2024. Large Language Models Are Not Robust Multiple Choice Selectors. *arXiv preprint arXiv:2309.03882v4*.

[letterpaper]article [submission]aaai2026 times helvet courier xcolor

Reproducibility Checklist

Instructions for Authors:

This document outlines key aspects for assessing reproducibility. Please provide your input by editing this `.tex` file directly.

For each question (that applies), replace the "Type your response here" text with your answer.

Example: If a question appears as

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
Type your response here
```

you would change it to:

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
yes
```

Please make sure to:

- Replace ONLY the "Type your response here" text and nothing else.
- Use one of the options listed for that question (e.g., **yes**, **no**, **partial**, or **NA**).
- **Not** modify any other part of the `\question` command or any other lines in this document.

You can `\input` this .tex file right before `\end{document}` of your main file or compile it as a stand-alone document. Check the instructions on your conference's website to see if you will be asked to provide this checklist with your paper or separately.

---

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) no

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) no

2.4. Proofs of all novel claims are included (yes/partial/no) no

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) no

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) no

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) NA

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) NA

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) yes

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) yes

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing

(yes/partial/no/NA) yes

(yes/partial/no/NA) yes

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) yes

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) yes

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) yes

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) yes

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) yes

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) yes

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) partial

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) yes

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) no

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) no

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments