Developing a Benchmark for 2D-to-3D Spatial Reasoning in Multimodal Large Language Models

Utkarsh Sharma, Ryan Spencer, Roey Yaari, Ritvik Vemavarapu, Joyce Yang, Steven Ngo

* Disclaimer: This is a research proposal. The final publication will be linked upon release. *

1. Introduction and Motivation

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 keeping track of object states. As AI systems interact more with real-world environments, 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; however, they often struggle with sequential spatial reasoning tasks that require understanding transformations over time, maintaining geometric consistency across viewpoints, or detecting physically impossible configurations. Existing benchmarks tend to evaluate static scenes, single-step reasoning, or end-state accuracy, leaving gaps in assessing procedural spatial understanding.

We propose the creation of a benchmark for sequential spatial planning and reasoning that uses origami-inspired tasks to couple 2D crease patterns with 3D multi-view states (front, back, top, bottom, left, right). Unlike prior work that primarily checks final answers in isolation, our benchmark will evaluate (1) multi-perspective consistency across 3D views, (2) feasibility via impossible-fold detection (violations of origami axioms), and (3) sequential interpretation of intermediate states induced by textual instructions and visual transitions.

2. Research Objectives

The primary objectives of this project are:

1. To develop a multi-view, sequential spatial reasoning benchmark that evaluates 2D-to-3D mapping in MLLMs.
2. To design evaluation metrics that capture nuanced failure modes, including:
    a. **Viewpoint Consistency:** Consistency of spatial predictions across multiple perspectives.
    b. **Angle-Sensitive Error:** Sensitivity to deviations in geometric transformations.
    c. **Impossible Fold Detection:** Ability to identify physically infeasible states.

3. To analyze model performance on tasks of varying complexity, thereby identifying strengths and weaknesses of existing MLLMs in procedural and geometric reasoning.

3. Methodology

3.1 Task Design

We will implement two clusters of tasks:

1. **Single-Step Spatial Understanding:** Models receive a 2D representation (e.g., a crease pattern) and predict the corresponding 3D structure.
2. **Multi-Step Spatial Reasoning:** We extend this reasoning over folding progressions, requiring models to be cognizant of geometric transitions and determine the step at which a fold can or cannot feasibly take place.

Tasks will vary in complexity, e.g., simple (<40 steps) vs. complex (≥40 steps), to examine the scaling of reasoning capabilities.

3.2 Dataset Construction

We will create a dataset of synthetic 2D-to-3D instances inspired by origami folding:

- Each instance includes a normalized 2D crease pattern representation and six 3D renderings from front, back, left, right, top, and bottom perspectives.
- Both physically valid and impossible transformations will be included to assess constraint awareness.

3.3 Evaluation Metrics

Performance will be measured using:

- **Accuracy** on final-state predictions.
- **Viewpoint Consistency** across six perspectives.
- **Angle-Sensitive Error** quantifying deviations from expected transformations.
- **Impossible Fold Selection Rate** to detect sensitivity to violations of physical constraints.

3.4 Model Evaluation

We will benchmark multiple open- and closed-source MLLMs to:

- Compare single-step vs. multi-step reasoning performance.
- Examine the effect of task complexity on spatial reasoning.
- Identify common failure modes, such as visual plausibility bias or inability to detect impossible configurations.

4. Expected Outcomes

- A publicly available benchmark for evaluating 2D-to-3D spatial reasoning in MLLMs.
- Insights into current MLLMs' strengths and weaknesses in procedural spatial tasks.
- A new diagnostic framework to guide the development of next-generation MLLMs, integrating geometry, perception, and physical reasoning.

5. Significance

This project addresses a key gap in multimodal AI research: assessing the ability of models to reason about space over time and across perspectives. By providing a structured benchmark, our work will enable systematic progress toward models that can plan, predict, and verify spatial transformations in real-world and simulated environments.

6. Timeline

- **Month 1:** Dataset design and 2D-to-3D instance generation.
- **Month 2:** Benchmark task implementation and evaluation metric development.
- **Month 3:** Model evaluation and analysis of failure modes.
- **Month 4:** Public release of the benchmark and initial research report.

7. Resources Required

- Computational resources for model inference and rendering 3D instances.
- Software frameworks for multimodal LLM evaluation.
- Expertise in 3D graphics, procedural generation, and AI benchmarking.