Novelty Assessment Report

Paper: Unfolding Spatial Cognition: Evaluating Multimodal Models on Visual Simulations
PDF URL: https://openreview.net/pdf?id=fbGmSV6tUw
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

Spatial cognition is essential for human intelligence, enabling problem-solving through visual simulations rather than solely relying on verbal reasoning. However, existing AI benchmarks primarily assess verbal reasoning, neglecting the complexities of non-verbal, multi-step visual simulation. We introduce \textbf{STARE (Spatial Transformations and Reasoning Evaluation)}, a benchmark designed to rigorously evaluate multimodal large language models on tasks better solved through multi-step visual simulation. STARE features 3K tasks spanning foundational geometric transformations (2D and 3D), integrated spatial reasoning (cube net folding and tangram puzzles), and real-world spatial reasoning (perspective and temporal reasoning), reflecting practical cognitive challenges like object assembly, mechanical diagram interpretation, and everyday spatial navigation. Our evaluations show that models excel at reasoning over simpler 2D transformations, but perform close to random chance on more complex tasks like 3D cube net folding and tangram puzzles that require multi-step visual simulations. Humans achieve near-perfect accuracy but take considerable time (up to 28.9s) on complex tasks, significantly speeding up (down by 7.5 seconds on average) with intermediate visual simulations. In contrast, models exhibit inconsistent performance gains from visual simulations, improving on most tasks but declining in specific cases like tangram puzzles (GPT-4o, o1) and cube net folding (Claude-3.5, Gemini-2.0 Flash), indicating that models may not know how to effectively leverage intermediate visual information.

Core Task Landscape

This paper addresses: **evaluating spatial reasoning through multi-step visual simulation**

A total of **50 papers** were analyzed and organized into a taxonomy with **18 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Spatial Reasoning Benchmarks and Evaluation Frameworks**
- **Spatial Reasoning Enhancement Methods for Vision-Language Models**
- **Application Domains for Spatial Reasoning**
- **Cognitive and Theoretical Foundations of Spatial Reasoning**
- **Auxiliary Studies and Methodological Tools**

Complete Taxonomy Tree

- evaluating spatial reasoning through multi-step visual simulation Survey Taxonomy
- Spatial Reasoning Benchmarks and Evaluation Frameworks
  - Multi-Step Visual Transformation and Simulation Tasks ★ (6 papers)
  - [0] Unfolding Spatial Cognition: Evaluating Multimodal Models on Visual Simulations (Anon et al., 2026) View paper
  - [5] Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the stepgame benchmark (Li, 2024) View paper
  - [12] LEGO-Puzzles: How Good Are MLLMs at Multi-Step Spatial Reasoning? (Gao Junyao, 2025) View paper
  - [14] VisionCube: 3D-Aware Vision-Language Model for Multi-Step Spatial Reasoning (Feiyang Wang, 2025) View paper
  - [34] Can World Simulators Reason? Gen-ViRe: A Generative Visual Reasoning Benchmark (Xinxin Liu, 2025) View paper
  - [35] ORIGAMISPACE: Benchmarking Multimodal LLMs in Multi-Step Spatial Reasoning with Mathematical Constraints (Rui Xu, 2025) View paper
  - Perspective-Taking and Viewpoint Transformation Evaluation (3 papers)
  - [3] Perspective-Aware Reasoning in Vision-Language Models via Mental Imagery Simulation (Phillip Y. Lee, 2025) View paper
  - [8] Spatial Mental Modeling from Limited Views (Yin, 2025) View paper
  - [50] SpinBench: Perspective and Rotation as a Lens on Spatial Reasoning in VLMs (Zhang Yu-you, 2025) View paper
  - Multi-Image and Cross-View Spatial Reasoning Assessment (2 papers)
  - [11] Seeing is Not Reasoning: MVPBench for Graph-based Evaluation of Multi-path Visual Physical CoT (Yi Junchao, 2025) View paper
  - [49] MMSI-Bench: A Benchmark for Multi-Image Spatial Intelligence (Yang Si-han, 2025) View paper
  - Real-World Simulation and Qualitative Spatial Reasoning Benchmarks (2 papers)
  - [31] Reframing Spatial Reasoning Evaluation in Language Models: A Real-World Simulation Benchmark for Qualitative Reasoning. (Li, 2024) View paper
  - [32] Benchmarking and enhancing spatial reasoning in large language models (Li, 2025) View paper
  - Domain-Specific Spatial Reasoning Evaluation (3 papers)
  - [22] Investigating the Impact of Spatial Reasoning on Construction Hazard Recognition (Seoyoung Cheon, 2025) View paper
  - [45] FRIEDA: Benchmarking Multi-Step Cartographic Reasoning in Vision-Language Models (Jiyoon Pyo, 2025) View paper
  - [47] Clinical trainee performance on taskâ□□based AR/VRâ□□guided surgical simulation is correlated with their 3D image spatial reasoning scores (Roy Eagleson, 2024) View paper

- Spatial Reasoning Enhancement Methods for Vision-Language Models
  - Visual Grounding and Spatial Attention Mechanisms (4 papers)
  - [1] Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing (Wu, 2025) View paper
  - [15] Grounded Reinforcement Learning for Visual Reasoning (Sarch, 2025) View paper
  - [16] Learning GUI Grounding with Spatial Reasoning from Visual Feedback (Zhao Yu, 2025) View paper
  - [25] High-Resolution Visual Reasoning via Multi-Turn Grounding-Based Reinforcement Learning (Huang XinYu, 2025) View paper
  - 3D-Aware Spatial Representation and Reconstruction (2 papers)
  - [2] Spatialrgpt: Grounded spatial reasoning in vision-language models (An-Chieh Cheng, 2024) View paper
  - [29] SpatialThinker: Reinforcing 3D Reasoning in Multimodal LLMs via Spatial Rewards (Hunar Batra, 2025) View paper
  - Structured Spatial Reasoning and Scene Graph Construction (3 papers)
  - [13] Hierarchical Spatial Proximity Reasoning for Vision-and-Language Navigation (Xu Ming, 2024) View paper
  - [27] Graphhopper: Multi-Hop Scene Graph Reasoning for Visual Question Answering (Koner, 2021) View paper
  - [36] Spatial Understanding from Videos: Structured Prompts Meet Simulation Data (Zhang Haoyu, 2025) View paper
  - Prompting Strategies and Chain-of-Thought for Spatial Reasoning (3 papers)
  - [38] LaV-CoT: Language-Aware Visual CoT with Multi-Aspect Reward Optimization for Real-World Multilingual VQA (Huang Jing, 2025) View paper
  - [42] LlamaV-o1: Rethinking Step-by-step Visual Reasoning in LLMs (Thawakar, 2025) View paper
  - [44] LLMs and Spatial Reasoning: Assessing Roadblocks and Providing Pathways to Improvement (William Peng, 2024) View paper
  - Knowledge Distillation and Data Generation for Spatial Reasoning (2 papers)
  - [17] AutoSpatial: Visual-Language Reasoning for Social Robot Navigation through Efficient Spatial Reasoning Learning (Kong, 2025) View paper
  - [40] SpatialTraceGen: High-Fidelity Traces for Efficient VLM Spatial Reasoning Distillation (Sheth, 2025) View paper
- Application Domains for Spatial Reasoning
  - Vision-and-Language Navigation and Instruction Following (4 papers)
  - [9] Exploring Spatial Representation to Enhance LLM Reasoning in Aerial Vision-Language Navigation (Gao Yunpeng, 2024) View paper
  - [18] Govig: Goal-conditioned visual navigation instruction generation (Wu Fengyi, 2025) View paper
  - [43] Aerial Vision-Language Navigation with a Unified Framework for Spatial, Temporal and Embodied Reasoning (Huilin Xu, 2025) View paper
  - [46] TransNav: spatial sequential transformer network for visual navigation (Kangjun Zhou, 2022) View paper
  - Robotic Manipulation and Action Planning (3 papers)
  - [6] MolmoAct: Action Reasoning Models that can Reason in Space (Lee, 2025) View paper
  - [7] AHA: A Vision-Language-Model for Detecting and Reasoning Over Failures in Robotic Manipulation (Duan, 2024) View paper
  - [33] Robotic Visual Instruction (Yanbang Li, 2025) View paper
  - Visual Question Answering with Spatial and Multi-Step Reasoning (3 papers)
  - [21] Dynamic Key-value Memory Enhanced Multi-step Graph Reasoning for Knowledge-based Visual Question Answering (LI Mingxiao, 2022) View paper
  - [30] Explore Multi-Step Reasoning in Video Question Answering (Yahong Han, 2018) View paper
  - [37] DVD: A Diagnostic Dataset for Multi-step Reasoning in Video Grounded Dialogue (Hung Le, 2021) View paper
  - Embodied AI and Interactive Environment Reasoning (2 papers)
  - [41] EmbRACE-3K: Embodied Reasoning and Action in Complex Environments (Lin Mingxian, 2025) View paper
  - [48] Spatial reasoning for real-time simulated environments (Nallacharu, 2005) View paper
  - Image Editing and Visual Content Generation with Spatial Reasoning (2 papers)
  - [4] Pixel-level reasoning segmentation via multi-turn conversations (Dunbo Cai, 2025) View paper
  - [10] Learning Action and Reasoning-Centric Image Editing from Videos and Simulations (Krojer, 2024) View paper
  - Specialized Domain Applications (2 papers)
  - [20] Large (vision) language models for autonomous vehicles: Current trends and future directions (Hanlin Tian, 2024) View paper
  - [28] Knowledge-Guided Large Language Models for Enhancing Agent-Based Wildfire Spatial Simulation (Ying Nie, 2025) View paper
- Cognitive and Theoretical Foundations of Spatial Reasoning (3 papers)
  - [19] Learned spatiotemporal sequence recognition and prediction in primary visual cortex (Jeffrey P. Gavornik, 2014) View paper
  - [24] A theory and a computational model of spatial reasoning with preferred mental models. (Marco Ragni, 2013) View paper
  - [39] Qualitative Spatial Reasoning and Simulation of Mechanismâ□□s Configuration (Zhang, 2005) View paper
- Auxiliary Studies and Methodological Tools (2 papers)
  - [23] Measurement of Spatial Contact Map Using Sequential FISH. (Hiroaki Ohishi, 2025) View paper
  - [26] Fiveâ□□Tails Multiâ□□Turn Visual Reasoning (Sato, 2025) View paper

Narrative

Core task: evaluating spatial reasoning through multi-step visual simulation. The field has organized itself around several complementary branches. The largest branch, Spatial Reasoning Benchmarks and Evaluation Frameworks, encompasses diverse diagnostic tasks—ranging from multi-step visual transformations like those in StepGame Benchmark[5] and LEGO Puzzles[12], to static spatial relation tests and dynamic simulation challenges such as VisionCube[14]. A second major branch, Spatial Reasoning Enhancement Methods for Vision-Language Models, explores techniques to improve model performance, including chain-of-thought prompting, mental imagery simulation approaches like Mental Imagery Simulation[3], and specialized architectural modifications. Application Domains for Spatial Reasoning demonstrates how these capabilities transfer to real-world settings—navigation, robotics, autonomous driving, and GUI interaction—while Cognitive and Theoretical Foundations of Spatial Reasoning draws on psychology and neuroscience to inform computational design. A smaller Auxiliary Studies branch provides methodological tools and cross-cutting analyses.

Within the benchmarking landscape, a particularly active line of work focuses on multi-step visual transformation and simulation tasks that require models to predict the outcome of sequential physical or geometric operations. Unfolding Spatial Cognition[0] sits squarely in this cluster, emphasizing iterative visual state changes that test whether models can mentally simulate unfolding processes. Nearby efforts like StepGame Benchmark[5] and VisionCube[14] similarly probe step-by-step reasoning but differ in their choice of domain—StepGame uses board-game-like scenarios while VisionCube targets 3D cube rotations. In contrast, Gen ViRe[34] and ORIGAMISPACE[35] explore generative or origami-specific transformations, highlighting trade-offs between procedural fidelity and task complexity. Across these works, open questions persist about the granularity of intermediate supervision, the role of explicit mental models versus end-to-end learning, and how well performance on synthetic benchmarks transfers to embodied or real-world spatial tasks.

Related Works in Same Category

The following **5 sibling papers** share the same taxonomy leaf node with the original paper:

1. Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the stepgame benchmark

Authors: Li, Fangjun, Hogg, David C., Fangjun Li, et al. (8 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Artificial intelligence (AI) has made remarkable progress across various domains, with large language models like ChatGPT gaining substantial attention for their human-like text-generation capabilities. Despite these achievements, improving spatial reasoning remains a significant challenge for these models. Benchmarks like StepGame evaluate AI spatial reasoning, where ChatGPT has shown unsatisfactory performance. However, the presence of template errors in the benchmark has an impact on the eval...

Relationship Analysis

Both papers belong to the Multi-Step Visual Transformation and Simulation Tasks category, evaluating spatial reasoning through sequential transformations. While the original paper (STARE) focuses on visual simulations across 2D/3D transformations, cube net folding, tangram puzzles, and real-world scenarios using multimodal models, the candidate paper (StepGame) concentrates on text-based multi-hop spatial reasoning in language models using natural language descriptions of directional relations. The key difference is that STARE emphasizes visual simulation capabilities with explicit intermediate visual steps, whereas StepGame evaluates purely linguistic spatial reasoning without visual components, using chain-of-thought and tree-of-thought prompting strategies to enhance LLM performance.

---

2. LEGO-Puzzles: How Good Are MLLMs at Multi-Step Spatial Reasoning?

Authors: Gao Junyao, Kexian Tang, Zeng Yan-hong, Junyao Gao, Duan, et al. (16 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Multi-step spatial reasoning entails understanding and reasoning about spatial relationships across multiple sequential steps, which is crucial for tackling complex real-world applications, such as robotic manipulation, autonomous navigation, and automated assembly. To assess how well current Multimodal Large Language Models (MLLMs) have acquired this fundamental capability, we introduce LEGO-Puzzles, a scalable benchmark designed to evaluate both spatial understanding and sequential reasoning i...

Relationship Analysis

Both papers belong to the Multi-Step Visual Transformation and Simulation Tasks category, evaluating spatial reasoning through sequential visual transformations and multi-step simulations. They overlap significantly in their focus on assessing MLLMs' abilities to perform multi-step spatial reasoning tasks, including 2D/3D transformations and puzzle-solving tasks like tangram and cube net folding, with both providing intermediate visual steps to test whether models can leverage explicit simulations. The key difference is that LEGO-Puzzles uses LEGO-based tasks as the primary evaluation framework and additionally explores image generation capabilities, while the original paper (STARE) employs a broader range of synthetic geometric transformations and real-world scenarios (perspective and temporal reasoning) without the LEGO constraint or generation component.

---

3. VisionCube: 3D-Aware Vision-Language Model for Multi-Step Spatial Reasoning

Authors: Feiyang Wang, Nan Luo, Wangyu Wu | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Solving a Rubik's Cube requires precise spatial reasoning, sequential planning, and adaptive decision-making. Traditional solvers depend on hand-crafted heuristics or symbolic planning, which limit generalizability across diverse cube states. In this work, we introduce VisionCube, a 3D-aware vision-language model that combines multi-view spatial reasoning with multimodal embodied planning to tackle Rubik's Cube manipulation. VisionCube integrates three core components: (1) Cube3D, which reconstr...

Relationship Analysis

Both papers belong to the Multi-Step Visual Transformation and Simulation Tasks category, evaluating spatial reasoning through sequential visual transformations requiring mental imagery. While the original paper (STARE) presents a comprehensive benchmark evaluating multimodal models across diverse spatial reasoning tasks (2D/3D transformations, cube net folding, tangram puzzles, perspective reasoning), the candidate paper (VisionCube) focuses specifically on solving Rubik's Cube manipulation through 3D-aware vision-language modeling with multi-view spatial reasoning and embodied planning. The key difference is that STARE is a broad evaluation benchmark testing models' visual simulation capabilities across multiple task types, whereas VisionCube is a specialized system designed for a single complex spatial task (Rubik's Cube solving) with explicit 3D reconstruction and robotic manipulation components.

---

4. Can World Simulators Reason? Gen-ViRe: A Generative Visual Reasoning Benchmark

Authors: Xinxin Liu, Zhaopan Xu, Ming Li, Kai Wang, Yong Jae Lee, et al. (6 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

While Chain-of-Thought (CoT) prompting enables sophisticated symbolic reasoning in LLMs, it remains confined to discrete text and cannot simulate the continuous, physics-governed dynamics of the real world. Recent video generation models have emerged as potential world simulators through Chain-of-Frames (CoF) reasoning -- materializing thought as frame-by-frame visual sequences, with each frame representing a physically-grounded reasoning step. Despite compelling demonstrations, a challenge pers...

Relationship Analysis

Both papers belong to the Multi-Step Visual Transformation and Simulation Tasks category, evaluating spatial reasoning through sequential visual processes. They overlap in assessing models' abilities to perform multi-step visual simulations (e.g., 2D/3D transformations, puzzle-solving) and both provide intermediate visual steps to test reasoning capabilities. However, the original paper (STARE) focuses on explicit step-by-step spatial transformations with controlled difficulty levels and human comparison studies, while the candidate paper (Gen-ViRe) emphasizes Chain-of-Frames reasoning in video generation models as world simulators, incorporating broader cognitive dimensions including planning, analogical reasoning, and real-world embodied AI scenarios.

---

5. ORIGAMISPACE: Benchmarking Multimodal LLMs in Multi-Step Spatial Reasoning with Mathematical Constraints

Authors: Rui Xu, Dakuan Lu, Zicheng Zhao, Xiaoyu Tan, Xintao Wang, et al. (8 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Spatial reasoning is a key capability in the field of artificial intelligence, especially crucial in areas such as robotics, computer vision, and natural language understanding. However, evaluating the ability of multimodal large language models(MLLMs) in complex spatial

reasoning still faces challenges, particularly in scenarios requiring multi-step reasoning and precise mathematical constraints. This paper introduces ORIGAMISPACE, a new dataset and benchmark designed to evaluate the multi-step...

Relationship Analysis

Both papers belong to the Multi-Step Visual Transformation and Simulation Tasks category, evaluating spatial reasoning through sequential visual transformations requiring mental imagery. They overlap in assessing models' abilities to perform multi-step spatial reasoning with explicit intermediate visual states (STARE provides intermediate visualizations for 2D/3D transformations, cube nets, and tangrams; ORIGAMISPACE provides folding process sequences). The key difference is that STARE focuses on diverse spatial tasks (geometric transformations, cube folding, tangrams, perspective reasoning) with synthesized data and human-like visual simulation evaluation, while ORIGAMISPACE specializes exclusively in origami tasks with strict mathematical constraints (Kawasaki's theorem, Huzita-Hatori axioms) and includes an interactive compiler environment for reinforcement learning-based code generation.

Contributions Analysis

Overall novelty summary. The paper introduces STARE, a benchmark evaluating spatial cognition through multi-step visual simulation across 2D/3D transformations, cube net folding, tangram puzzles, and real-world scenarios. It resides in the 'Multi-Step Visual Transformation and Simulation Tasks' leaf, which contains six papers including the original work. This leaf sits within the broader 'Spatial Reasoning Benchmarks and Evaluation Frameworks' branch, indicating a moderately populated research direction focused specifically on sequential geometric manipulations requiring mental imagery rather than single-step or navigation-based tasks.

The taxonomy reveals neighboring evaluation approaches: 'Perspective-Taking and Viewpoint Transformation Evaluation' (three papers on mental rotation and viewpoint shifts), 'Multi-Image and Cross-View Spatial Reasoning Assessment' (two papers on 3D inference from multiple views), and 'Real-World Simulation and Qualitative Spatial Reasoning Benchmarks' (two papers on realistic 3D scenarios). STARE's emphasis on integrated tasks like tangram puzzles and cube net folding bridges abstract geometric transformations with practical assembly challenges, positioning it between purely synthetic benchmarks and domain-specific evaluations like cartography or construction safety assessments found in adjacent leaves.

Among thirty candidates examined, none clearly refute the three core contributions: the STARE benchmark itself (ten candidates, zero refutable), the evaluation framework comparing reasoning with and without intermediate visual simulations (ten candidates, zero refutable), and the comprehensive analysis of model limitations (ten candidates, zero refutable). The sibling papers in the same leaf—StepGame Benchmark, VisionCube, and others—address related sequential reasoning but differ in domain focus (board games, 3D rotations) or task granularity, suggesting STARE's combination of foundational transformations with integrated puzzles occupies a distinct niche within this limited search scope.

Based on the top-thirty semantic matches and taxonomy structure, STARE appears to contribute a novel task suite blending geometric primitives with complex assembly challenges. The absence of refutable prior work in this limited search does not guarantee exhaustive novelty but indicates that among closely related benchmarks examined, none directly anticipate STARE's specific combination of 2D/3D transformations, tangram puzzles, and cube net folding with explicit visual simulation evaluation.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: STARE benchmark for evaluating spatial cognition through visual simulations

Description: The authors introduce STARE, a comprehensive benchmark containing approximately 4,000 tasks spanning foundational geometric transformations (2D and 3D), integrated spatial reasoning (cube net folding and tangram puzzles), and real-world spatial reasoning (perspective and temporal reasoning). The benchmark is specifically designed to evaluate whether multimodal models can perform complex visual reasoning through multi-step visual simulations, similar to how humans solve spatial problems.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Mind the gap: Benchmarking spatial reasoning in vision-language models

URL: View paper

Brief Assessment

Mind the Gap[52] focuses on evaluating spatial reasoning components (spatial relations, orientation, mental rotation, spatial visualization) using static images and psychometric tests, whereas the original paper's STARE benchmark specifically evaluates multi-step visual simulations (e.g., cube net folding, tangram puzzles) with intermediate visual states. The candidate does not demonstrate prior work on benchmarking visual simulation capabilities.

---

2. SPATIA: Multimodal Model for Prediction and Generation of Spatial Cell Phenotypes

URL: View paper

Brief Assessment

SPATIA[59] focuses on spatial transcriptomics in biology, integrating cell morphology and gene expression data. It does not address spatial reasoning or visual simulation benchmarks for multimodal AI models.

---

3. Rank2Tell: A Multimodal Driving Dataset for Joint Importance Ranking and Reasoning

URL: View paper

Brief Assessment

Rank2Tell[58] focuses on autonomous driving scene understanding with importance ranking and reasoning about traffic objects, not on evaluating spatial cognition or visual simulation capabilities in multimodal models.

---

4. 11plus-bench: Demystifying multimodal llm spatial reasoning with cognitive-inspired analysis

URL: View paper

Brief Assessment

11plus Bench[57] focuses on standardized spatial aptitude tests with cognitive feature annotations for human-model comparison, whereas STARE emphasizes multi-step visual simulation tasks (2D/3D transformations, cube folding, tangram). The benchmarks target different evaluation paradigms and do not challenge STARE's novelty in visual simulation-based assessment.

---

5. Multi-modal learning for geospatial vegetation forecasting

URL: View paper

Brief Assessment

Geospatial Vegetation Forecasting[56] focuses on predicting vegetation greenness from satellite imagery using multi-modal transformers, not on evaluating spatial reasoning or visual simulation capabilities in multimodal models through benchmark tasks.

---

6. Thinking in space: How multimodal large language models see, remember, and recall spaces

URL: View paper

Brief Assessment

Thinking in Space[53] focuses on video-based spatial memory and recall tasks (e.g., object counting, distance estimation from videos), not on evaluating multi-step visual simulations of geometric transformations like cube net folding or tangram puzzles that STARE emphasizes.

---

7. Govig: Goal-conditioned visual navigation instruction generation

URL: View paper

Brief Assessment

Govig[18] focuses on goal-conditioned visual navigation instruction generation from egocentric observations, not on evaluating spatial cognition or visual simulation capabilities in multimodal models through benchmark tasks.

---

8. What is the visual cognition gap between humans and multimodal llms?

URL: View paper

Brief Assessment

Visual Cognition Gap[55] focuses on matrix reasoning tasks (e.g., Raven's Progressive Matrices) to evaluate abstract pattern recognition and rule inference, whereas the original paper targets spatial transformations and visual simulations (2D/3D transformations, cube net folding, tangram puzzles). These are distinct cognitive domains with different task structures and evaluation goals.

---

9. Visfactor: Benchmarking fundamental visual cognition in multimodal large language models

URL: View paper

Brief Assessment

Visfactor[54] focuses on digitizing standardized human cognitive assessments (FRCT battery) to evaluate foundational visual cognition factors like closure, memory, and perceptual speed. STARE specifically targets multi-step visual simulation capabilities in spatial reasoning tasks with explicit intermediate visualizations, representing a distinct evaluation approach.

---

10. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models

URL: View paper

Brief Assessment

Visulogic[51] focuses on visual logical reasoning across diverse categories (quantitative, spatial, positional, attribute, stylistic) without emphasizing multi-step visual simulations or providing intermediate simulation states, which are central to STARE's design.

Contribution 2: Evaluation framework with and without intermediate visual simulations

Description: The authors develop a systematic evaluation framework that tests models under different conditions: with only questions, with textual step descriptions, and with explicit intermediate visual simulations. This framework enables fine-grained analysis of whether models can effectively leverage visual guidance versus relying solely on internal mental simulation capabilities.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. EscapeCraft: A 3D Room Escape Environment for Benchmarking Complex Multimodal Reasoning Ability

URL: View paper

Brief Assessment

EscapeCraft[65] focuses on room escape tasks in 3D environments with action-based interactions, not on spatial transformation tasks with explicit intermediate visual simulations for geometric reasoning.

---

2. Spatial Mental Modeling from Limited Views

URL: View paper

Brief Assessment

Spatial Mental Modeling[8] focuses on spatial mental models for scene understanding from limited views, not on evaluating models with/without intermediate visual simulations for spatial reasoning tasks like the original paper's framework.

---

3. Perspective-Aware Reasoning in Vision-Language Models via Mental Imagery Simulation

URL: View paper

Brief Assessment

Mental Imagery Simulation[3] focuses on perspective-aware reasoning through abstract scene representations for viewpoint transformation, not on evaluating models with/without intermediate visual simulations for spatial reasoning tasks.

---

4. Reframing Spatial Reasoning Evaluation in Language Models: A Real-World Simulation Benchmark for Qualitative Reasoning.

URL: View paper

Brief Assessment

Real World Simulation[31] focuses on qualitative spatial reasoning in realistic 3D environments without providing intermediate visual simulations. The ORIGINAL paper's framework explicitly tests models with and without step-by-step visual guidance for spatial transformations, which is not addressed in this candidate.

---

5. SpatialRGPT: Grounded Spatial Reasoning in Vision Language Model

URL: View paper

Brief Assessment

SpatialRGPT Grounded[61] focuses on enhancing spatial perception through regional representations and depth integration for robotic tasks, not on evaluating models with/without intermediate visual simulations for spatial reasoning.

---

6. VisualTrans: A Benchmark for Real-World Visual Transformation Reasoning

URL: View paper

Brief Assessment

VisualTrans[63] focuses on real-world human-object interaction scenarios with manipulation tasks, not on systematic evaluation of models' ability to leverage intermediate visual simulations versus internal mental simulation across diverse spatial reasoning tasks.

---

7. SSR: Enhancing Depth Perception in Vision-Language Models via Rationale-Guided Spatial Reasoning

URL: View paper

Brief Assessment

SSR Depth Perception[62] focuses on depth-based spatial reasoning with textual rationales for depth perception tasks, not on evaluating models with/without intermediate visual simulations across diverse spatial transformation tasks.

---

8. Learning to Localize Objects Improves Spatial Reasoning in Visual-LLMs

URL: View paper

Brief Assessment

Localize Objects Reasoning[64] focuses on injecting spatial awareness into visual-LLMs through coordinate-based instruction fine-tuning, not on evaluating models with/without intermediate visual simulations for spatial reasoning tasks.

---

9. Scene-LLM: Extending Language Model for 3D Visual Understanding and Reasoning

URL: View paper

Brief Assessment

Scene LLM[60] focuses on 3D visual understanding in interactive indoor environments using hybrid 3D feature representations, not on evaluation frameworks that systematically test models with and without intermediate visual simulations for spatial reasoning tasks.

---

10. Clinical trainee performance on taskâ□□based AR/VRâ□□guided surgical simulation is correlated with their 3D image spatial reasoning scores

URL: View paper

Brief Assessment

The candidate paper focuses on AR/VR surgical training evaluation using spatial reasoning tests for neurosurgical procedures, not on evaluating multimodal AI models with intermediate visual simulations for spatial reasoning tasks.

Contribution 3: Comprehensive analysis revealing model limitations in visual simulation

Description: The authors provide extensive experimental analysis demonstrating that current multimodal models struggle with complex spatial reasoning tasks requiring multi-step visual simulations, performing near random chance on tasks like cube net folding and tangram puzzles. They reveal that models exhibit inconsistent performance gains from visual simulations and identify specific failure modes including perception errors and inability to integrate visual context effectively.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. LEGO-Puzzles: How Good Are MLLMs at Multi-Step Spatial Reasoning?

URL: View paper

Brief Assessment

LEGO Puzzles[12] focuses on LEGO-based spatial understanding and sequential reasoning tasks, while the original paper examines multi-step visual simulation across geometric transformations, cube net folding, and tangram puzzles. The task domains and evaluation methodologies differ substantially.

---

2. SpatialReasoner: Towards Explicit and Generalizable 3D Spatial Reasoning

URL: View paper

Brief Assessment

SpatialReasoner[71] focuses on 3D spatial reasoning with explicit 3D representations (locations, orientations) rather than evaluating multi-step visual simulation capabilities across diverse spatial transformation tasks like cube net folding and tangram puzzles that the original paper studies.

---

3. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark

URL: View paper

Brief Assessment

EMMA Benchmark[66] focuses on multimodal reasoning across math, physics, chemistry, and coding domains, while the original paper specifically targets multi-step visual simulation tasks like cube net folding and tangram puzzles. The candidate does not demonstrate prior work on the specific spatial reasoning tasks and visual simulation challenges that form the core of the original contribution.

---

4. ORIGAMISPACE: Benchmarking Multimodal LLMs in Multi-Step Spatial Reasoning with Mathematical Constraints

URL: View paper

Brief Assessment

ORIGAMISPACE[35] focuses on origami-based spatial reasoning tasks with mathematical constraints, not general multi-step visual simulation across diverse spatial reasoning domains like cube net folding and tangram puzzles evaluated in the original paper.

---

5. Open3dvqa: A benchmark for comprehensive spatial reasoning with multimodal large language model in open space

URL: View paper

Brief Assessment

Open3DVQA[69] focuses on spatial reasoning in 3D urban environments from aerial perspectives, evaluating distance, direction, and size reasoning tasks. The original paper examines multi-step visual simulations for tasks like cube net folding and tangram puzzles, which are fundamentally different spatial reasoning challenges.

---

6. Thinking in space: How multimodal large language models see, remember, and recall spaces

URL: View paper

Brief Assessment

Thinking in Space[53] analyzes spatial reasoning from videos but does not examine model performance on multi-step visual simulations with intermediate steps, which is the core focus of the original paper's analysis.

---

7. Visual-o1: Understanding ambiguous instructions via multi-modal multi-turn chain-of-thoughts reasoning

URL: View paper

Brief Assessment

Visual-o1[70] focuses on understanding ambiguous language instructions in multi-modal tasks through chain-of-thought reasoning, not on evaluating multi-step visual simulation capabilities for spatial reasoning tasks like cube net folding and tangram puzzles.

---

8. High-Resolution Visual Reasoning via Multi-Turn Grounding-Based Reinforcement Learning

URL: View paper

Brief Assessment

Multi Turn Grounding[25] focuses on enabling models to iteratively crop and focus on key image regions in high-resolution scenarios through reinforcement learning, rather than analyzing model limitations in multi-step visual simulation for spatial reasoning tasks like cube net folding and tangram puzzles.

---

9. Multi-modal situated reasoning in 3d scenes

URL: View paper

Brief Assessment

Multi Modal 3D[68] focuses on situated reasoning in 3D scenes with multi-modal inputs (text, images, point clouds), not on multi-step visual simulation tasks like cube net folding or tangram puzzles that require sequential mental transformations.

---

10. Vocot: Unleashing visually grounded multi-step reasoning in large multi-modal models

URL: View paper

Brief Assessment

VoCoT[67] focuses on developing a multi-step reasoning framework for LMMs using object-centric chain-of-thought, not on analyzing model limitations in visual simulation for spatial reasoning tasks like cube folding or tangram puzzles.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Unfolding Spatial Cognition: Evaluating Multimodal Models on Visual Simulations View paper
- [1] Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing View paper
- [2] Spatialrgpt: Grounded spatial reasoning in vision-language models View paper
- [3] Perspective-Aware Reasoning in Vision-Language Models via Mental Imagery Simulation View paper
- [4] Pixel-level reasoning segmentation via multi-turn conversations View paper
- [5] Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the stepgame benchmark View paper
- [6] MolmoAct: Action Reasoning Models that can Reason in Space View paper
- [7] AHA: A Vision-Language-Model for Detecting and Reasoning Over Failures in Robotic Manipulation View paper
- [8] Spatial Mental Modeling from Limited Views View paper
- [9] Exploring Spatial Representation to Enhance LLM Reasoning in Aerial Vision-Language Navigation View paper
- [10] Learning Action and Reasoning-Centric Image Editing from Videos and Simulations View paper
- [11] Seeing is Not Reasoning: MVPBench for Graph-based Evaluation of Multi-path Visual Physical CoT View paper
- [12] LEGO-Puzzles: How Good Are MLLMs at Multi-Step Spatial Reasoning? View paper
- [13] Hierarchical Spatial Proximity Reasoning for Vision-and-Language Navigation View paper
- [14] VisionCube: 3D-Aware Vision-Language Model for Multi-Step Spatial Reasoning View paper
- [15] Grounded Reinforcement Learning for Visual Reasoning View paper
- [16] Learning GUI Grounding with Spatial Reasoning from Visual Feedback View paper
- [17] AutoSpatial: Visual-Language Reasoning for Social Robot Navigation through Efficient Spatial Reasoning Learning View paper
- [18] Govig: Goal-conditioned visual navigation instruction generation View paper
- [19] Learned spatiotemporal sequence recognition and prediction in primary visual cortex View paper
- [20] Large (vision) language models for autonomous vehicles: Current trends and future directions View paper
- [21] Dynamic Key-value Memory Enhanced Multi-step Graph Reasoning for Knowledge-based Visual Question Answering View paper
- [22] Investigating the Impact of Spatial Reasoning on Construction Hazard Recognition View paper
- [23] Measurement of Spatial Contact Map Using Sequential FISH. View paper
- [24] A theory and a computational model of spatial reasoning with preferred mental models. View paper
- [25] High-Resolution Visual Reasoning via Multi-Turn Grounding-Based Reinforcement Learning View paper
- [26] Fiveâ□□Tails Multiâ□□Turn Visual Reasoning View paper
- [27] Graphhopper: Multi-Hop Scene Graph Reasoning for Visual Question Answering View paper
- [28] Knowledge-Guided Large Language Models for Enhancing Agent-Based Wildfire Spatial Simulation View paper
- [29] SpatialThinker: Reinforcing 3D Reasoning in Multimodal LLMs via Spatial Rewards View paper
- [30] Explore Multi-Step Reasoning in Video Question Answering View paper
- [31] Reframing Spatial Reasoning Evaluation in Language Models: A Real-World Simulation Benchmark for Qualitative Reasoning. View paper
- [32] Benchmarking and enhancing spatial reasoning in large language models View paper
- [33] Robotic Visual Instruction View paper
- [34] Can World Simulators Reason? Gen-ViRe: A Generative Visual Reasoning Benchmark View paper
- [35] ORIGAMISPACE: Benchmarking Multimodal LLMs in Multi-Step Spatial Reasoning with Mathematical Constraints View paper
- [36] Spatial Understanding from Videos: Structured Prompts Meet Simulation Data View paper
- [37] DVD: A Diagnostic Dataset for Multi-step Reasoning in Video Grounded Dialogue View paper
- [38] LaV-CoT: Language-Aware Visual CoT with Multi-Aspect Reward Optimization for Real-World Multilingual VQA View paper
- [39] Qualitative Spatial Reasoning and Simulation of Mechanismâ□□s Configuration View paper
- [40] SpatialTraceGen: High-Fidelity Traces for Efficient VLM Spatial Reasoning Distillation View paper
- [41] EmbRACE-3K: Embodied Reasoning and Action in Complex Environments View paper
- [42] LlamaV-o1: Rethinking Step-by-step Visual Reasoning in LLMs View paper
- [43] Aerial Vision-Language Navigation with a Unified Framework for Spatial, Temporal and Embodied Reasoning View paper

- [44] LLMs and Spatial Reasoning: Assessing Roadblocks and Providing Pathways to Improvement View paper
- [45] FRIEDA: Benchmarking Multi-Step Cartographic Reasoning in Vision-Language Models View paper
- [46] TransNav: spatial sequential transformer network for visual navigation View paper
- [47] Clinical trainee performance on taskâ□□based AR/VRâ□□guided surgical simulation is correlated with their 3D image spatial reasoning scores View paper
- [48] Spatial reasoning for real-time simulated environments View paper
- [49] MMSI-Bench: A Benchmark for Multi-Image Spatial Intelligence View paper
- [50] SpinBench: Perspective and Rotation as a Lens on Spatial Reasoning in VLMs View paper
- [51] Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models View paper
- [52] Mind the gap: Benchmarking spatial reasoning in vision-language models View paper
- [53] Thinking in space: How multimodal large language models see, remember, and recall spaces View paper
- [54] Visfactor: Benchmarking fundamental visual cognition in multimodal large language models View paper
- [55] What is the visual cognition gap between humans and multimodal llms? View paper
- [56] Multi-modal learning for geospatial vegetation forecasting View paper
- [57] 11plus-bench: Demystifying multimodal llm spatial reasoning with cognitive-inspired analysis View paper
- [58] Rank2Tell: A Multimodal Driving Dataset for Joint Importance Ranking and Reasoning View paper
- [59] SPATIA: Multimodal Model for Prediction and Generation of Spatial Cell Phenotypes View paper
- [60] Scene-LLM: Extending Language Model for 3D Visual Understanding and Reasoning View paper
- [61] SpatialRGPT: Grounded Spatial Reasoning in Vision Language Model View paper
- [62] SSR: Enhancing Depth Perception in Vision-Language Models via Rationale-Guided Spatial Reasoning View paper
- [63] VisualTrans: A Benchmark for Real-World Visual Transformation Reasoning View paper
- [64] Learning to Localize Objects Improves Spatial Reasoning in Visual-LLMs View paper
- [65] EscapeCraft: A 3D Room Escape Environment for Benchmarking Complex Multimodal Reasoning Ability View paper
- [66] Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark View paper
- [67] Vocot: Unleashing visually grounded multi-step reasoning in large multi-modal models View paper
- [68] Multi-modal situated reasoning in 3d scenes View paper
- [69] Open3dvqa: A benchmark for comprehensive spatial reasoning with multimodal large language model in open space View paper
- [70] Visual-o1: Understanding ambiguous instructions via multi-modal multi-turn chain-of-thoughts reasoning View paper
- [71] SpatialReasoner: Towards Explicit and Generalizable 3D Spatial Reasoning View paper