Novelty Assessment Report

Paper: Seeing Across Views: Benchmarking Spatial Reasoning of Vision-Language Models in Robotic Scenes

PDF URL: https://openreview.net/pdf?id=jXDZJAfRZB

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-01

Abstract

Vision-language models (VLMs) are essential to Embodied AI, enabling robots to perceive, reason, and act in complex environments. They also serve as the foundation for the recent Vision-Language-Action (VLA) models. Yet, most evaluations of VLMs focus on single-view settings, leaving their ability to integrate multi-view information largely underexplored. At the same time, multi-camera setups are increasingly standard in robotic platforms, as they provide complementary perspectives to mitigate occlusion and depth ambiguity. Whether VLMs can effectively leverage such multi-view inputs for robotic reasoning therefore remains an open question. To bridge this gap, we introduce MV-RoboBench, a benchmark specifically designed to evaluate the multi-view spatial reasoning capabilities of VLMs in robotic manipulation. MV-RoboBench consists of 1.7k manually curated QA items across eight subtasks, divided into two primary categories: spatial understanding and robotic execution. We evaluate a diverse set of existing VLMs, including both open-source and closed-source models, along with enhanced versions augmented by Chain-of-Thought (CoT)-inspired enhancements. The results show that state-of-the-art models remain far below human performance, underscoring the substantial challenges VLMs face in multi-view robotic perception. Additionally, our analysis uncovers two key findings: (i) spatial intelligence and robotic task reasoning are correlated in multi-view robotic scenarios; and (ii) strong performance on existing general-purpose single-view spatial understanding benchmarks does not reliably translate to success in the robotic spatial tasks assessed by our benchmark. We release MV-RoboBench as an open resource to foster progress in spatially grounded VLMs and VLAs, providing a foundation for advancing embodied multi-view intelligence in robotics.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Multi-View Spatial Reasoning in Robotic Manipulation**

A total of **50 papers** were analyzed and organized into a taxonomy with **23 categories**.

Taxonomy Overview

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

- **Multi-View 3D Representation Learning**
- **Vision-Language Integration for Spatial Reasoning**
- **Spatial Reasoning Mechanisms and Representations**
- **Action Representation and Execution**
- **End-to-End Policy Learning**
- **Relational and Graph-Based Reasoning**
- **Data Generation and Augmentation**
- **Interactive and Closed-Loop Reasoning**
- **Spatial Reasoning for Human-Robot Collaboration**
- **Benchmarking and Evaluation**
- ... and 1 more categories

Complete Taxonomy Tree

- Multi-View Spatial Reasoning in Robotic Manipulation Survey Taxonomy
- Multi-View 3D Representation Learning
  - Transformer-Based Multi-View Encoding (4 papers)
  - [2] 3d-mvp: 3d multiview pretraining for robotic manipulation (Qian, 2024) View paper
  - [3] Rvt: Robotic view transformer for 3d object manipulation (Goyal Ankit, 2023) View paper
  - [11] 3D-MVP: 3D Multiview Pretraining for Manipulation (Shengyi Qian, 2025) View paper
  - [21] Act3D: 3D Feature Field Transformers for Multi-Task Robotic Manipulation (Gervet, 2023) View paper
  - Multi-View Pretraining and Self-Supervised Learning (2 papers)
  - [4] CL3R: 3D Reconstruction and Contrastive Learning for Enhanced Robotic Manipulation Representations (Cui Wenbo, 2025) View paper
  - [24] Robouniview: Visual-language model with unified view representation for robotic manipulation (Liu Fan-fan, 2024) View paper
  - Temporal and Historical Multi-View Integration (3 papers)
  - [15] Temporal consistent multi-view perception for robust embodied manipulation (Haoyuan Chen, 2025) View paper
  - [20] Integrating Historical Learning and Multi-View Attention with Hierarchical Feature Fusion for Robotic Manipulation (Gaoxiong Lu, 2024) View paper
  - [44] Addressing Information Loss in Multi-View Robotic Manipulation Through History Observation (Shuyu Zhang, 2025) View paper
  - Multi-View Fusion for Scene Understanding (2 papers)
  - [25] Multi-view fusion for multi-level robotic scene understanding (Lin Yunzhi, 2021) View paper
  - [36] Towards Global Awareness in Human-Robot-Collaborative Multi-cell Assembly System (Khansa Rekik, 2024) View paper

- Vision-Language Integration for Spatial Reasoning
  - VLM-Based Spatial Understanding and Grounding (3 papers)
  - [6] Incentivizing Multimodal Reasoning in Large Models for Direct Robot Manipulation (Tang Wei-liang, 2025) View paper
  - [8] RoboRefer: Towards Spatial Referring with Reasoning in Vision-Language Models for Robotics (Zhou En-shen, 2025) View paper
  - [23] RoboSpatial: Teaching Spatial Understanding to 2D and 3D Vision-Language Models for Robotics (Chan-Hee Song, 2024) View paper
  - Spatial Referring and Language-Grounded Localization (2 papers)
  - [16] Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation (Qi, 2025) View paper
  - [42] Spatial Reasoning from Natural Language Instructions for Robot Manipulation (Venkatesh, 2021) View paper
  - Chain-of-Thought and Hierarchical Reasoning (2 papers)
  - [13] SpatialCoT: Advancing Spatial Reasoning through Coordinate Alignment and Chain-of-Thought for Embodied Task Planning (Liu Yue-cheng, 2025) View paper
  - [28] VisionCube: 3D-Aware Vision-Language Model for Multi-Step Spatial Reasoning (Feiyang Wang, 2025) View paper
- Spatial Reasoning Mechanisms and Representations
  - Explicit Spatial Encoding and Grid-Based Methods (2 papers)
  - [10] Grid-augmented vision: A simple yet effective approach for enhanced spatial understanding in multi-modal agents (Joongwon Chae, 2024) View paper
  - [12] GP3: A 3D Geometry-Aware Policy with Multi-View Images for Robotic Manipulation (Zhao Guo-yang, 2025) View paper
  - Orientation Understanding and Semantic Alignment (1 papers)
  - [17] Right side up? disentangling orientation understanding in mllms with fine-grained multi-axis perception tasks (Tasnim, 2025) View paper
  - Cross-View Correspondence and Point Matching (2 papers)
  - [38] Towards Cross-View Point Correspondence in Vision-Language Models (Yipu Wang, 2025) View paper
  - [49] Multi-View Registration of Partially Overlapping Point Clouds for Robotic Manipulation (Yuzhen Xie, 2024) View paper
  - Geometry-Aware Policy Learning (1 papers)
  - [43] SpatialActor: Exploring Disentangled Spatial Representations for Robust Robotic Manipulation (Hao Shi, 2025) View paper
- Action Representation and Execution
  - Intermediate Representation for Action Grounding (3 papers)
  - [5] Embodied-r1: Reinforced embodied reasoning for general robotic manipulation (Yuan, 2025) View paper
  - [14] From Seeing to Doing: Bridging Reasoning and Decision for Robotic Manipulation (Yuan, 2025) View paper
  - [18] Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints (Mingjie Pan, 2025) View paper
  - Diffusion-Based Action Generation (2 papers)
  - [31] PDFactor: Learning Tri-Perspective View Policy Diffusion Field for Multi-Task Robotic Manipulation (Jingyi Tian, 2025) View paper
  - [41] Spatial Policy: Guiding Visuomotor Robotic Manipulation with Spatial-Aware Modeling and Reasoning (Liu Yi-jun, 2025) View paper
  - Visual Kinematic Chain and Embodiment-Agnostic Actions (1 papers)
  - [45] Scaling manipulation learning with visual kinematic chain prediction (Zhang Xin-yu, 2024) View paper
- End-to-End Policy Learning
  - Imitation Learning with Visual Foundation Models (2 papers)
  - [26] CLIPort: What and Where Pathways for Robotic Manipulation (Mohit Shridhar, 2021) View paper
  - [27] SAM-E: Leveraging Visual Foundation Model with Sequence Imitation for Embodied Manipulation (Zhang Junâ□□Jie, 2024) View paper
  - Reinforcement Learning for Manipulation (2 papers)
  - [29] Learning to Manipulate Anywhere: A Visual Generalizable Framework For Reinforcement Learning (Yuan, 2024) View paper
  - [47] Deep Reinforcement Learning-Based Robotic Grasping in Clutter and Occlusion (Marwan Qaid Mohammed, 2021) View paper
  - Multi-Task and Long-Horizon Policy Learning (2 papers)
  - [9] Enerverse: Envisioning embodied future space for robotics manipulation (Huang, 2025) View paper
  - [50] RoboHorizon: An LLM-Assisted Multi-View World Model for Long-Horizon Robotic Manipulation (Chen ZiXuan, 2025) View paper
- Relational and Graph-Based Reasoning (2 papers)
  - [19] Dual Graph Attention Networks for Multi-View Visual Manipulation Relationship Detection and Robotic Grasping (Mengyuan Ding, 2025) View paper
  - [46] Spatial Reasoning via Deep Vision Models for Robotic Sequential Manipulation (Hongyou Zhou, 2023) View paper
- Data Generation and Augmentation (2 papers)
  - [1] Robo2vlm: Visual question answering from large-scale in-the-wild robot manipulation datasets (Chen, 2025) View paper
  - [30] ERMV: Editing 4D Robotic Multi-view images to enhance embodied agents (Nie Chang, 2025) View paper
- Interactive and Closed-Loop Reasoning (2 papers)
  - [39] Closed Loop Interactive Embodied Reasoning for Robot Manipulation (Michal Nazarczuk, 2025) View paper
  - [40] Using meta-reasoning for incremental repairs in multi-object robot manipulation tasks (Priyam Parashar, 2022) View paper
- Spatial Reasoning for Human-Robot Collaboration (3 papers)
  - [7] Cognitive reasoning for compliant robot manipulation (Daniel Sebastian Leidner, 2019) View paper
  - [35] Spatial reasoning for human robot interaction (E. A. Sisbot, 2007) View paper
  - [48] A Multi-view Framework for Human Parsing in Human-Robot Collaboration scenarios (Terreran, 2021) View paper
- Benchmarking and Evaluation ★ (2 papers)
  - [0] Seeing Across Views: Benchmarking Spatial Reasoning of Vision-Language Models in Robotic Scenes (Anon et al., 2026) View paper
  - [37] MineAnyBuild: Benchmarking Spatial Planning for Open-world AI Agents (Wei Ziming, 2025) View paper
- Theoretical Foundations and Survey Works (4 papers)
  - [22] Visual reasoning for robotics manipulation (Nazarczuk, 2024) View paper
  - [32] Embodied intelligence for robot manipulation: development and challenges (Honghao Song, 2025) View paper
  - [33] Deep 3D Geometric Reasoning for Robot Manipulation (Eisner, 2025) View paper

◦ [34] Spatial reasoning for real-time robotic manipulation (Hanyoung Jang, 2006) View paper

Narrative

Core task: multi-view spatial reasoning in robotic manipulation scenarios. The field organizes itself around several complementary branches that address different facets of enabling robots to understand and act upon spatial information from multiple viewpoints. Multi-View 3D Representation Learning focuses on how to encode geometric structure from camera arrays, with works like Robotic View Transformer[3] and 3D Multiview Pretraining[2] building rich volumetric or feature-based representations. Vision-Language Integration for Spatial Reasoning explores how natural language instructions can guide spatial understanding, exemplified by approaches such as Robo2vlm[1] and RoboRefer[8]. Spatial Reasoning Mechanisms and Representations delve into explicit reasoning modules—ranging from coordinate transforms to attention over spatial features—while Action Representation and Execution and End-to-End Policy Learning address how spatial understanding translates into executable robot actions. Relational and Graph-Based Reasoning captures methods that model object interactions and scene structure explicitly, and Data Generation and Augmentation tackles the challenge of obtaining diverse training scenarios. Interactive and Closed-Loop Reasoning, Spatial Reasoning for Human-Robot Collaboration, and Benchmarking and Evaluation round out the taxonomy by addressing online adaptation, collaborative settings, and systematic assessment of spatial reasoning capabilities.

Recent activity highlights a tension between end-to-end learned policies and modular pipelines that separate perception, reasoning, and control. Many studies pursue tighter integration of vision and language to handle complex instructions, as seen in Embodied-r1[5] and Incentivizing Multimodal Reasoning[6], while others emphasize robust 3D scene understanding through multi-view fusion or explicit spatial representations like SpatialCoT[13] and RoboSpatial[23]. Seeing Across Views[0] sits squarely within the Benchmarking and Evaluation branch, providing systematic assessment tools for multi-view spatial reasoning rather than proposing a new architecture. Its emphasis contrasts with neighboring work like MineAnyBuild[37], which also evaluates spatial capabilities but does so in a construction-focused interactive environment. By offering structured benchmarks, Seeing Across Views[0] complements the broader ecosystem: it helps quantify progress across the diverse methodological branches and identifies which spatial reasoning challenges remain open, thereby guiding future work in representation learning, policy design, and human-robot collaboration.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. MineAnyBuild: Benchmarking Spatial Planning for Open-world AI Agents

Authors: Wei Ziming, Lin, Bingqian, Ziming Wei, Bingqian Lin, et al. (17 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Spatial Planning is a crucial part in the field of spatial intelligence, which requires the understanding and planning about object arrangements in space perspective. AI agents with the spatial planning ability can better adapt to various real-world applications, including robotic manipulation, automatic assembly, urban planning etc. Recent works have attempted to construct benchmarks for evaluating the spatial intelligence of Multimodal Large Language Models (MLLMs). Nevertheless, these benchma...

Relationship Analysis

Both papers belong to the Benchmarking and Evaluation category, focusing on assessing spatial reasoning capabilities in embodied AI contexts. While the original paper (MV-RoboBench) evaluates multi-view spatial reasoning in real robotic manipulation scenarios using synchronized camera views from physical robot demonstrations, MineAnyBuild evaluates spatial planning abilities of AI agents in the Minecraft virtual environment through architecture building tasks. The key difference lies in their evaluation domains: MV-RoboBench emphasizes multi-view perception integration for physical robotic tasks, whereas MineAnyBuild focuses on open-world spatial planning and creativity in a simulated game environment.

Contributions Analysis

Overall novelty summary. The paper introduces MV-RoboBench, a benchmark comprising 1.7k manually curated question-answer items designed to evaluate multi-view spatial reasoning capabilities of vision-language models in robotic manipulation contexts. Within the taxonomy, it resides in the 'Benchmarking and Evaluation' leaf, which contains only two papers total. This makes it a relatively sparse research direction compared to more crowded areas like 'Transformer-Based Multi-View Encoding' (four papers) or 'VLM-Based Spatial Understanding and Grounding' (three papers). The benchmark focuses on assessing existing models rather than proposing new architectures or learning methods.

The taxonomy reveals that most neighboring work concentrates on algorithmic contributions: multi-view representation learning, vision-language integration, and end-to-end policy learning. The 'Benchmarking and Evaluation' leaf sits apart from these methodological branches, serving a complementary role by providing systematic assessment tools. Its sibling paper in the same leaf addresses evaluation in a different context (construction-focused interactive environments), while nearby branches like 'Vision-Language Integration for Spatial Reasoning' and 'Multi-View 3D Representation Learning' develop the techniques that benchmarks like MV-RoboBench aim to measure. This positioning suggests the paper addresses a gap in systematic evaluation infrastructure rather than competing directly with method-oriented work.

Among 30 candidates examined through semantic search, none were found to clearly refute any of the three identified contributions: the MV-RoboBench benchmark itself (10 candidates examined, 0 refutable), the comprehensive VLM evaluation with CoT enhancements (10 candidates, 0 refutable), and the correlation analysis between spatial and robotic reasoning (10 candidates, 0 refutable). The absence of refutable prior work across all contributions suggests that within this limited search scope, the specific combination of multi-view focus, robotic manipulation context, and systematic VLM evaluation appears relatively unexplored. However, this reflects the scale of the search rather than an exhaustive literature review.

Based on the limited search of 30 semantically similar papers, the work appears to occupy a distinct position by providing evaluation infrastructure for multi-view spatial reasoning in robotics. The sparse 'Benchmarking and Evaluation' category and lack of overlapping prior work within the examined candidates suggest novelty in this specific assessment focus. However, the analysis does not cover the full breadth of vision-language or robotic benchmarking literature, and a more comprehensive search might reveal additional related evaluation efforts or datasets addressing similar capabilities.

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

Contribution 1: MV-RoboBench benchmark for multi-view spatial reasoning in robotic manipulation

Description: The authors present MV-RoboBench, the first benchmark that integrates spatial understanding and robotic execution tasks using synchronized multi-view camera inputs from real robotic demonstrations. It contains 1.7K manually curated QA items across eight subtasks to systematically evaluate vision-language models in multi-view robotic scenarios.

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[72] focuses on multi-step spatial reasoning using LEGO-based tasks without multi-view camera inputs or robotic execution components. The candidate evaluates spatial understanding through sequential LEGO assembly tasks, whereas the original paper specifically targets synchronized multi-view perception from real robotic demonstrations with camera arrays.

2. Multimodal spatial reasoning in the large model era: A survey and benchmarks

URL: View paper

Brief Assessment

Multimodal Spatial Survey[75] is a survey paper that reviews existing work in multimodal spatial reasoning. It does not present a benchmark for multi-view spatial reasoning in robotic manipulation, and therefore cannot refute the novelty of MV-RoboBench.

3. SITE: towards Spatial Intelligence Thorough Evaluation

URL: View paper

Brief Assessment

SITE[74] focuses on general spatial intelligence evaluation across diverse visual modalities (single-image, multi-image, video) and cognitive factors, not specifically on synchronized multi-view robotic manipulation scenarios with real robotic demonstrations as in MV-RoboBench.

4. SpaceVista: All-Scale Visual Spatial Reasoning from mm to km

URL: View paper

Brief Assessment

SpaceVista[71] focuses on all-scale spatial reasoning across diverse scenarios (mm to km) including robotics and autonomous driving, but does not specifically target multi-view spatial reasoning with synchronized camera inputs in robotic manipulation scenarios. The candidate addresses broader spatial scales rather than the specific multi-view robotic manipulation benchmark proposed in the original paper.

5. RoboMIND: Benchmark on Multi-embodiment Intelligence Normative Data for Robot Manipulation

URL: View paper

Brief Assessment

RoboMIND[73] is a dataset of 107k demonstration trajectories for robot manipulation, not a benchmark for evaluating multi-view spatial reasoning capabilities of vision-language models. The two works serve fundamentally different purposes in the robotics domain.

6. Rvt: Robotic view transformer for 3d object manipulation

URL: View paper

Brief Assessment

Robotic View Transformer[3] focuses on a multi-view transformer architecture for 3D object manipulation, not on creating a benchmark for evaluating vision-language models' spatial reasoning capabilities in multi-view robotic scenarios.

7. Act3D: 3D Feature Field Transformers for Multi-Task Robotic Manipulation

URL: View paper

Brief Assessment

Act3D[21] focuses on a manipulation policy architecture using 3D feature fields for action prediction, not on benchmarking multi-view spatial reasoning capabilities of vision-language models. The paper does not present any benchmark or evaluation framework for assessing VLM spatial understanding.

8. RoboHorizon: An LLM-Assisted Multi-View World Model for Long-Horizon Robotic Manipulation

URL: View paper

Brief Assessment

RoboHorizon[50] focuses on model-based visual RL for long-horizon manipulation tasks using multi-view world models, not on creating benchmarks for evaluating spatial reasoning capabilities of vision-language models in multi-view robotic scenarios.

9. RoboRefer: Towards Spatial Referring with Reasoning in Vision-Language Models for Robotics

URL: View paper

Brief Assessment

RoboRefer[8] focuses on spatial referring tasks with single-view depth integration and multi-step reasoning, introducing RefSpatial-Bench for evaluating spatial referring. This differs from MV-RoboBench's emphasis on synchronized multi-view camera inputs for spatial understanding and robotic execution tasks in manipulation scenarios.

10. ViewSpatial-Bench: Evaluating Multi-perspective Spatial Localization in Vision-Language Models

URL: View paper

Brief Assessment

ViewSpatial-Bench[70] focuses on multi-viewpoint spatial localization from different entity perspectives (egocentric vs. allocentric reasoning), not on robotic manipulation tasks with synchronized multi-camera inputs from real robotic demonstrations.

Contribution 2: Comprehensive evaluation of VLMs with CoT-inspired enhancements

Description: The authors conduct extensive experiments evaluating various VLMs on multi-view robotic reasoning and explore three CoT-inspired enhancement directions: textual scene descriptions, view synthesis, and depth priors. Their results show state-of-the-art models remain far below human performance.

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. Zero-shot object navigation with vision-language models reasoning

URL: View paper

Brief Assessment

Zero-shot Object Navigation[51] focuses on robot navigation and object interaction tasks, not comprehensive evaluation of VLMs on multi-view robotic reasoning with CoT-inspired enhancements across spatial understanding and robotic execution subtasks.

2. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation

URL: View paper

Brief Assessment

ManipLLM[53] focuses on robotic manipulation using multimodal LLMs with chain-of-thought reasoning for pose prediction, not on evaluating VLMs across multi-view robotic reasoning tasks or benchmarking spatial understanding capabilities.

3. VCoT-Grasp: Grasp Foundation Models with Visual Chain-of-Thought Reasoning for Language-driven Grasp Generation

URL: View paper

Brief Assessment

VCoT-Grasp[55] focuses on language-driven robotic grasping with visual chain-of-thought reasoning for grasp generation in manipulation tasks, not on comprehensive evaluation of VLMs across multi-view robotic reasoning benchmarks with CoT-inspired enhancements.

4. Robotic control via embodied chain-of-thought reasoning

URL: View paper

Brief Assessment

Embodied Chain-of-Thought[52] focuses on training vision-language-action models with embodied reasoning for robotic control, not on evaluating existing VLMs with CoT techniques across multi-view spatial reasoning tasks as in the original paper.

5. Generate subgoal images before act: Unlocking the chain-of-thought reasoning in diffusion model for robot manipulation with multimodal prompts

URL: View paper

Brief Assessment

Generate Subgoal Images[58] focuses on generating visual subgoal images for robot manipulation using diffusion models, not on evaluating VLMs with CoT techniques for multi-view robotic reasoning as in the original paper.

6. Embodiedgpt: Vision-language pre-training via embodied chain of thought

URL: View paper

Brief Assessment

EmbodiedGPT[56] focuses on embodied planning and control in robotics using chain-of-thought for generating step-by-step action plans, not on evaluating VLMs' multi-view spatial reasoning capabilities as in the original paper.

7. Embodied Chain of Action Reasoning with Multi-Modal Foundation Model for Humanoid Loco-manipulation

URL: View paper

Brief Assessment

Chain Action Reasoning[60] focuses on humanoid loco-manipulation tasks with chain-of-thought reasoning for robotic control, not on evaluating vision-language models across multi-view robotic reasoning benchmarks with CoT-inspired enhancements like textual descriptions, view synthesis, and depth priors.

8. Graphcot-vla: A 3d spatial-aware reasoning vision-language-action model for robotic manipulation with ambiguous instructions

URL: View paper

Brief Assessment

GraphCoT-VLA[59] proposes a structured chain-of-thought reasoning module for robotic manipulation with ambiguous instructions, focusing on 3D spatial reasoning and action planning. The original paper evaluates existing VLMs with CoT-inspired enhancements (textual descriptions, view synthesis, depth priors) on multi-view robotic reasoning benchmarks. These are distinct research contributions with different scopes and methodologies.

9. Emma-x: An embodied multimodal action model with grounded chain of thought and look-ahead spatial reasoning

URL: View paper

Brief Assessment

Emma-x[57] focuses on training a vision-language-action model for robotic manipulation with grounded chain-of-thought reasoning and spatial planning. It does not evaluate multiple existing VLMs on multi-view robotic reasoning or explore CoT-inspired enhancement directions like textual scene descriptions, view synthesis, and depth priors as the original paper does.

10. dvla: Diffusion vision-language-action model with multimodal chain-of-thought

URL: View paper

Brief Assessment

dvla[54] focuses on a unified diffusion-based vision-language-action model for robotic control, not on evaluating existing VLMs with Chain-of-Thought techniques across multi-view robotic reasoning tasks as in the original paper.

Contribution 3: Correlation analysis revealing relationships between spatial and robotic reasoning

Description: The authors provide systematic correlation analysis demonstrating that spatial and robotic reasoning are related in multi-view manipulation settings, while also showing that single-view spatial benchmark performance does not transfer reliably to multi-view robotic tasks, highlighting unique challenges of embodied multi-view intelligence.

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. Enhancing computational thinking and spatial reasoning skills in gamification programming learning: A comparative study of tangible, block and paperâ□□andâ□□pencil â□¦

URL: View paper

Brief Assessment

Gamification Programming Learning[61] focuses on educational programming environments with tangible/block-based interfaces for computational thinking, not on vision-language models' spatial reasoning capabilities in robotic manipulation scenarios.

2. Towards Unobtrusive Physical AI: Augmenting Everyday Objects with Intelligence and Robotic Movement for Proactive Assistance

URL: View paper

Brief Assessment

Unobtrusive Physical AI[64] focuses on proactive object assistance using LLMs for spatial-temporal reasoning in everyday tasks, not on systematic correlation analysis between spatial intelligence and robotic task reasoning capabilities in multi-view manipulation settings.

3. Robix: A unified model for robot interaction, reasoning and planning

URL: View paper

Brief Assessment

Robix[66] focuses on unified robot interaction and task planning with chain-of-thought reasoning for embodied tasks. It does not present correlation analysis between spatial intelligence and robotic reasoning capabilities, which is the core novelty claim of the original paper's contribution.

4. Enhancing computational thinking, Spatial reasoning, and executive function skills: The impact of tangible programming tools in early childhood and across different â¦

URL: View paper

Brief Assessment

Tangible Programming Tools[62] focuses on early childhood education using tangible programming tools to enhance computational thinking and spatial reasoning. It does not address multi-view robotic manipulation or vision-language models, making it irrelevant to the original paper's contribution on correlation between spatial intelligence and robotic task reasoning in embodied AI systems.

5. Unravelling the Computational Thinking and Spatial Thinking Development: An Exploration of a Virtual Robot Programming Environment

URL: View paper

Brief Assessment

Virtual Robot Programming[69] examines correlations between computational thinking (CT) and spatial thinking (ST) in educational robot programming contexts, not the relationship between spatial intelligence and robotic task reasoning in vision-language models for manipulation. The domains, methodologies, and research questions are fundamentally different.

6. Magma: A Foundation Model for Multimodal AI Agents

URL: View paper

Brief Assessment

Magma[63] focuses on training a foundation model for multimodal AI agents with action grounding capabilities (SOM/TOM), not on analyzing correlations between spatial and robotic reasoning abilities in multi-view settings.

7. Boosting Robotic Manipulation Generalization with Minimal Costly Data

URL: View paper

Brief Assessment

Minimal Costly Data[68] focuses on decoupling manipulation trajectories into spatial reasoning and physical interaction phases to improve data efficiency in VLA training. It does not provide systematic correlation analysis between spatial intelligence and robotic task reasoning capabilities across different model architectures or benchmark settings.

8. MolmoAct: Action Reasoning Models that can Reason in Space

URL: View paper

Brief Assessment

MolmoAct[65] focuses on action reasoning models that integrate perception, planning, and control for robotic manipulation. It does not provide correlation analysis between spatial intelligence and robotic task reasoning capabilities in multi-view settings.

9. From Seeing to Doing: Bridging Reasoning and Decision for Robotic Manipulation

URL: View paper

Brief Assessment

Seeing to Doing[14] focuses on developing a vision-language model for robotic manipulation through spatial relationship reasoning, but does not present systematic correlation analysis between spatial intelligence and robotic task reasoning capabilities as a research contribution.

10. Structured Task Solving via Modular Embodied Intelligence: A Case Study on Rubik's Cube

URL: View paper

Brief Assessment

Modular Embodied Intelligence[67] focuses on symbolic task planning for Rubik's cube manipulation using knowledge bases and LLMs, not on evaluating correlation between spatial intelligence and robotic reasoning capabilities in multi-view settings.

Appendix: Text Similarity Detection

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

References

- [0] Seeing Across Views: Benchmarking Spatial Reasoning of Vision-Language Models in Robotic Scenes View paper
- [1] Robo2vlm: Visual question answering from large-scale in-the-wild robot manipulation datasets View paper
- [2] 3d-mvp: 3d multiview pretraining for robotic manipulation View paper
- [3] Rvt: Robotic view transformer for 3d object manipulation View paper
- [4] CL3R: 3D Reconstruction and Contrastive Learning for Enhanced Robotic Manipulation Representations View paper
- [5] Embodied-r1: Reinforced embodied reasoning for general robotic manipulation View paper
- [6] Incentivizing Multimodal Reasoning in Large Models for Direct Robot Manipulation View paper
- [7] Cognitive reasoning for compliant robot manipulation View paper
- [8] RoboRefer: Towards Spatial Referring with Reasoning in Vision-Language Models for Robotics View paper
- [9] Enerverse: Envisioning embodied future space for robotics manipulation View paper

- [10] Grid-augmented vision: A simple yet effective approach for enhanced spatial understanding in multi-modal agents View paper
- [11] 3D-MVP: 3D Multiview Pretraining for Manipulation View paper
- [12] GP3: A 3D Geometry-Aware Policy with Multi-View Images for Robotic Manipulation View paper
- [13] SpatialCoT: Advancing Spatial Reasoning through Coordinate Alignment and Chain-of-Thought for Embodied Task Planning View paper
- [14] From Seeing to Doing: Bridging Reasoning and Decision for Robotic Manipulation View paper
- [15] Temporal consistent multi-view perception for robust embodied manipulation View paper
- [16] Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation View paper
- [17] Right side up? disentangling orientation understanding in mllms with fine-grained multi-axis perception tasks View paper
- [18] Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints View paper
- [19] Dual Graph Attention Networks for Multi-View Visual Manipulation Relationship Detection and Robotic Grasping View paper
- [20] Integrating Historical Learning and Multi-View Attention with Hierarchical Feature Fusion for Robotic Manipulation View paper
- [21] Act3D: 3D Feature Field Transformers for Multi-Task Robotic Manipulation View paper
- [22] Visual reasoning for robotics manipulation View paper
- [23] RoboSpatial: Teaching Spatial Understanding to 2D and 3D Vision-Language Models for Robotics View paper
- [24] Robouniview: Visual-language model with unified view representation for robotic manipulation View paper
- [25] Multi-view fusion for multi-level robotic scene understanding View paper
- [26] CLIPort: What and Where Pathways for Robotic Manipulation View paper
- [27] SAM-E: Leveraging Visual Foundation Model with Sequence Imitation for Embodied Manipulation View paper
- [28] VisionCube: 3D-Aware Vision-Language Model for Multi-Step Spatial Reasoning View paper
- [29] Learning to Manipulate Anywhere: A Visual Generalizable Framework For Reinforcement Learning View paper
- [30] ERMV: Editing 4D Robotic Multi-view images to enhance embodied agents View paper
- [31] PDFactor: Learning Tri-Perspective View Policy Diffusion Field for Multi-Task Robotic Manipulation View paper
- [32] Embodied intelligence for robot manipulation: development and challenges View paper
- [33] Deep 3D Geometric Reasoning for Robot Manipulation View paper
- [34] Spatial reasoning for real-time robotic manipulation View paper
- [35] Spatial reasoning for human robot interaction View paper
- [36] Towards Global Awareness in Human-Robot-Collaborative Multi-cell Assembly System View paper
- [37] MineAnyBuild: Benchmarking Spatial Planning for Open-world AI Agents View paper
- [38] Towards Cross-View Point Correspondence in Vision-Language Models View paper
- [39] Closed Loop Interactive Embodied Reasoning for Robot Manipulation View paper
- [40] Using meta-reasoning for incremental repairs in multi-object robot manipulation tasks View paper
- [41] Spatial Policy: Guiding Visuomotor Robotic Manipulation with Spatial-Aware Modeling and Reasoning View paper
- [42] Spatial Reasoning from Natural Language Instructions for Robot Manipulation View paper
- [43] SpatialActor: Exploring Disentangled Spatial Representations for Robust Robotic Manipulation View paper
- [44] Addressing Information Loss in Multi-View Robotic Manipulation Through History Observation View paper
- [45] Scaling manipulation learning with visual kinematic chain prediction View paper
- [46] Spatial Reasoning via Deep Vision Models for Robotic Sequential Manipulation View paper
- [47] Deep Reinforcement Learning-Based Robotic Grasping in Clutter and Occlusion View paper
- [48] A Multi-view Framework for Human Parsing in Human-Robot Collaboration scenarios View paper
- [49] Multi-View Registration of Partially Overlapping Point Clouds for Robotic Manipulation View paper
- [50] RoboHorizon: An LLM-Assisted Multi-View World Model for Long-Horizon Robotic Manipulation View paper
- [51] Zero-shot object navigation with vision-language models reasoning View paper
- [52] Robotic control via embodied chain-of-thought reasoning View paper
- [53] Manipllm: Embodied multimodal large language model for object-centric robotic manipulation View paper
- [54] dvla: Diffusion vision-language-action model with multimodal chain-of-thought View paper
- [55] VCoT-Grasp: Grasp Foundation Models with Visual Chain-of-Thought Reasoning for Language-driven Grasp Generation View paper
- [56] Embodiedgpt: Vision-language pre-training via embodied chain of thought View paper
- [57] Emma-x: An embodied multimodal action model with grounded chain of thought and look-ahead spatial reasoning View paper
- [58] Generate subgoal images before act: Unlocking the chain-of-thought reasoning in diffusion model for robot manipulation with multimodal prompts View paper
- [59] Graphcot-vla: A 3d spatial-aware reasoning vision-language-action model for robotic manipulation with ambiguous instructions View paper
- [60] Embodied Chain of Action Reasoning with Multi-Modal Foundation Model for Humanoid Loco-manipulation View paper
- [61] Enhancing computational thinking and spatial reasoning skills in gamification programming learning: A comparative study of tangible, block and paperâ□□andâ□□pencil â□¦ View paper
- [62] Enhancing computational thinking, Spatial reasoning, and executive function skills: The impact of tangible programming tools in early childhood and across different â□¦ View paper
- [63] Magma: A Foundation Model for Multimodal AI Agents View paper
- [64] Towards Unobtrusive Physical AI: Augmenting Everyday Objects with Intelligence and Robotic Movement for Proactive Assistance View paper
- [65] MolmoAct: Action Reasoning Models that can Reason in Space View paper
- [66] Robix: A unified model for robot interaction, reasoning and planning View paper
- [67] Structured Task Solving via Modular Embodied Intelligence: A Case Study on Rubik's Cube View paper
- [68] Boosting Robotic Manipulation Generalization with Minimal Costly Data View paper
- [69] Unravelling the Computational Thinking and Spatial Thinking Development: An Exploration of a Virtual Robot Programming Environment View paper
- [70] ViewSpatial-Bench: Evaluating Multi-perspective Spatial Localization in Vision-Language Models View paper
- [71] SpaceVista: All-Scale Visual Spatial Reasoning from mm to km View paper
- [72] LEGO-Puzzles: How Good Are MLLMs at Multi-Step Spatial Reasoning? View paper
- [73] RoboMIND: Benchmark on Multi-embodiment Intelligence Normative Data for Robot Manipulation View paper
- [74] SITE: towards Spatial Intelligence Thorough Evaluation View paper
- [75] Multimodal spatial reasoning in the large model era: A survey and benchmarks View paper