Novelty Assessment Report

Paper: MomaGraph: State-Aware Unified Scene Graphs with Vision-Language Models for Embodied Task Planning
PDF URL: https://openreview.net/pdf?id=3eTr9dGwJv
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

Mobile manipulators in households must both navigate and manipulate. This requires a compact, semantically rich scene representation that captures where objects are, how they function, and which parts are actionable. Scene graphs are a natural choice, yet prior work often separates spatial and functional relations, treats scenes as static snapshots without object states or temporal updates, and overlooks information most relevant for accomplishing the current task. To overcome these shortcomings, we introduce MomaGraph, a unified scene representation for embodied agents that integrates spatial-functional relationships and part-level interactive elements. However, advancing such a representation requires both suitable data and rigorous evaluation, which have been largely missing. To address this, we construct MomaGraph-Scenes, the first large-scale dataset of richly annotated, task-driven scene graphs in household environments, and design MomaGraph-Bench, a systematic evaluation suite spanning six reasoning capabilities from high-level planning to fine-grained scene understanding. Built upon this foundation, we further develop MomaGraph-R1, a 7B vision–language model trained with reinforcement learning on MomaGraph-Scenes. MomaGraph-R1 predicts task-oriented scene graphs and serves as a zero-shot task planner under a Graph-then-Plan framework. Extensive experiments show that our model achieves state-of-the-art results among open-source models, reaching 71.6% accuracy on the benchmark (+11.4% over the best baseline), while generalizing across public benchmarks and transferring effectively to real-robot experiments. More visualizations and robot demonstrations are available at https://momagraph.github.io/.

Core Task Landscape

This paper addresses: **Task-Oriented Scene Graph Generation for Embodied Agents**

A total of **50 papers** were analyzed and organized into a taxonomy with **20 categories**.

Taxonomy Overview

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

- **Scene Graph Construction and Representation**
- **Task Planning and Reasoning with Scene Graphs**
- **Navigation with Scene Graphs**
- **Manipulation and Interaction with Scene Graphs**
- **Scene Generation and Simulation**
- **Domain-Specific Applications**
- **Representation Learning and Continuous Embeddings**
- **Benchmarking and Datasets**

Complete Taxonomy Tree

- Task-Oriented Scene Graph Generation for Embodied Agents Survey Taxonomy
- Scene Graph Construction and Representation
  - 3D Metric-Semantic Scene Graph Generation (4 papers)
  - [1] Open-vocabulary functional 3d scene graphs for real-world indoor spaces (Chenyangguang Zhang, 2025) View paper
  - [3] Grapheqa: Using 3d semantic scene graphs for real-time embodied question answering (Saxena, 2024) View paper
  - [11] TACS-Graphs: Traversability-Aware Consistent Scene Graphs for Ground Robot Indoor Localization and Mapping (Kim Jeewon, 2025) View paper
  - [38] 3D dynamic scene graphs: Actionable spatial perception with places, objects, and humans (Rosinol, 2020) View paper
  - Functional and Interactive Scene Graphs (4 papers)
  - [2] Roboexp: Action-conditioned scene graph via interactive exploration for robotic manipulation (Jiang, 2024) View paper
  - [4] VLM-MSGraph: Vision Language Model-enabled Multi-hierarchical Scene Graph for robotic assembly (Shufei Li, 2025) View paper
  - [23] Dynamic Interactive Relation Capturing via Scene Graph Learning for Robotic Surgical Report Generation (Hongqiu Wang, 2023) View paper
  - [24] Hi-Dyna Graph: Hierarchical Dynamic Scene Graph for Robotic Autonomy in Human-Centric Environments (Hou Jia-wei, 2025) View paper
  - Dynamic and Temporal Scene Graph Updating (3 papers)
  - [32] Time is on my sight: scene graph filtering for dynamic environment perception in an LLM-driven robot (Simone Colombani, 2024) View paper
  - [37] Lost & Found: Tracking Changes From Egocentric Observations in 3D Dynamic Scene Graphs (Tjark Behrens, 2024) View paper
  - [39] Belief Scene Graphs: Expanding Partial Scenes with Objects through Computation of Expectation (Mario A. V. Saucedo, 2024) View paper
  - Open-Vocabulary and Foundation Model-Based Scene Graphs (2 papers)

- [46] Open scene graphs for open-world object-goal navigation (Joel Loo, 2025) View paper
- [48] Mapping High-level Semantic Regions in Indoor Environments without Object Recognition (Roberto Bigazzi, 2024) View paper

- Task Planning and Reasoning with Scene Graphs
  - LLM-Based Task Planning with Scene Graphs ★ (5 papers)
  - [0] MomaGraph: State-Aware Unified Scene Graphs with Vision-Language Models for Embodied Task Planning (Anon et al., 2026) View paper
  - [9] SayPlan: Grounding Large Language Models using 3D Scene Graphs for Scalable Task Planning (Rana, 2023) View paper
  - [27] EmbodiedRAG: Dynamic 3D Scene Graph Retrieval for Efficient and Scalable Robot Task Planning (Meghan Booker, 2024) View paper
  - [28] Hierarchical Generation of Action Sequence for Service Rots Based on Scene Graph via Large Language Models (Yu Gu, 2024) View paper
  - [30] Context Matters! Relaxing Goals with LLMs for Feasible 3D Scene Planning (Brienza, 2025) View paper
  - Schema-Guided and Multi-Agent Reasoning (2 papers)
  - [16] Schema-Guided Scene-Graph Reasoning based on Multi-Agent Large Language Model System (Chen, 2025) View paper
  - [26] Information-Theoretic Graph Fusion with Vision-Language-Action Model for Policy Reasoning and Dual Robotic Control (Shunlei Li, 2025) View paper
  - Classical and Symbolic Planning with Scene Graphs (2 papers)
  - [13] Long-term robot manipulation task planning with scene graph and semantic knowledge (Runqing Miao, 2023) View paper
  - [20] Reasoning with scene graphs for robot planning under partial observability (Saeid Amiri, 2022) View paper
  - Memory-Augmented and Context-Aware Planning (3 papers)
  - [35] SituationalLLM: Proactive Language Models with Scene Awareness for Dynamic, Contextual Task Guidance (Muhammad Saif Ullah Khan, 2024) View paper
  - [40] Orionnav: Online planning for robot autonomy with context-aware llm and open-vocabulary semantic scene graphs (Goswami, 2024) View paper
  - [45] KARMA: Augmenting Embodied AI Agents with Long-and-Short Term Memory Systems (Zixuan Wang, 2024) View paper
- Navigation with Scene Graphs
  - Object-Goal Navigation with Scene Graphs (4 papers)
  - [5] Commonsense scene graph-based target localization for object search (Wenqi Ge, 2024) View paper
  - [19] SG-Nav: Online 3D Scene Graph Prompting for LLM-based Zero-shot Object Navigation (Yin Hang, 2024) View paper
  - [36] Task-Driven Graph Attention for Hierarchical Relational Object Navigation (Michael Lingelbach, 2023) View paper
  - [42] SOON: Scenario Oriented Object Navigation with Graph-based Exploration (Fengda Zhu, 2021) View paper
  - Vision-Language Navigation with Scene Graphs (4 papers)
  - [7] Learning Bird's Eye View scene graph and knowledge-inspired policy for embodied visual navigation (Jian Luo, 2025) View paper
  - [12] Scene Graph Contrastive Learning for Embodied Navigation (Kunal Pratap Singh, 2023) View paper
  - [31] General Scene Adaptation for Vision-and-Language Navigation (Qiao, 2025) View paper
  - [49] Multi-modal scene graph inspired policy for visual navigation (Yu He, 2024) View paper
  - Spatial Reasoning and Geospatial Navigation (2 papers)
  - [18] EmbodiedVSR: Dynamic Scene Graph-Guided Chain-of-Thought Reasoning for Visual Spatial Tasks (Zhang Yi, 2025) View paper
  - [34] GeoNav: Empowering MLLMs with Explicit Geospatial Reasoning Abilities for Language-Goal Aerial Navigation (Xu, 2025) View paper
  - Multi-Modal and Knowledge-Driven Navigation (2 papers)
  - [21] Generation of skill-specific maps from graph world models for robotic systems (DE Vos Koen, 2024) View paper
  - [50] Knowledge-driven Scene Priors for Semantic Audio-Visual Embodied Navigation (Tatiya, 2022) View paper
- Manipulation and Interaction with Scene Graphs
  - Object Rearrangement and Spatial Manipulation (1 papers)
  - [41] SG-Bot: Object Rearrangement via Coarse-to-Fine Robotic Imagination on Scene Graphs (Guangyao Zhai, 2023) View paper
  - Dynamics and Interaction Prediction (2 papers)
  - [33] GraphAD: Interaction Scene Graph for End-to-end Autonomous Driving (Zhang Yunpeng, 2024) View paper
  - [43] Graph-based Task-specific Prediction Models for Interactions between Deformable and Rigid Objects (Zehang Weng, 2021) View paper
- Scene Generation and Simulation
  - Task-Driven and Interactive Scene Generation (3 papers)
  - [6] Mesatask: Towards task-driven tabletop scene generation via 3d spatial reasoning (Jinkun Hao, 2025) View paper
  - [10] Agentworld: An interactive simulation platform for scene construction and mobile robotic manipulation (Zhang Yizheng, 2025) View paper
  - [14] Dynamic Scene Generation for Embodied Navigation Benchmark (C Wang, 2024) View paper
  - Text-to-3D Scene Synthesis with Foundation Models (2 papers)
  - [22] Scenethesis: A language and vision agentic framework for 3d scene generation (Ling Lu, 2025) View paper
  - [29] Agentic 3D Scene Generation with Spatially Contextualized VLMs (Liu Xinhang, 2025) View paper
  - Scene Augmentation and Diversity Enhancement (1 papers)
  - [17] Scene Augmentation Methods for Interactive Embodied AI Tasks (Hongrui Sang, 2023) View paper
- Domain-Specific Applications
  - Surgical Robotics and Medical Applications (1 papers)
  - [47] VISAGE: Video Synthesis using Action Graphs for Surgery (Yeganeh, 2024) View paper
- Representation Learning and Continuous Embeddings (1 papers)
  - [44] Continuous Scene Representations for Embodied AI (Samir Yitzhak Gadre, 2022) View paper
- Benchmarking and Datasets (3 papers)
  - [8] In defense of scene graph generation for human-robot open-ended interaction in service robotics (Maelic Neau, 2023) View paper
  - [15] 3D scene graphs in robotics: A unified representation bridging geometry, semantics, and action (Iacopo Catalano, 2025) View paper
  - [25] Scene-Driven Multimodal Knowledge Graph Construction for Embodied AI (Yaoxian Song, 2023) View paper

Narrative

Core task: task-oriented scene graph generation for embodied agents. The field centers on building structured, graph-based representations of environments that enable robots and virtual agents to reason about objects, spatial relationships, and affordances in service of concrete tasks. The taxonomy reveals several major branches: Scene Graph Construction and Representation focuses on how to extract and maintain these graphs from sensor data, often leveraging vision-language models or open-vocabulary methods like Open Vocabulary Functional Graphs[1]. Task Planning and Reasoning with Scene Graphs explores how agents use these structures to decompose high-level goals into executable steps, with many studies integrating large language models to ground symbolic reasoning in spatial context. Navigation with Scene Graphs addresses how agents exploit relational cues for efficient exploration and goal-directed movement, while Manipulation and Interaction branches examine grasping, tool use, and dynamic updates to the graph as the agent acts. Additional branches cover Scene Generation and Simulation for synthetic training data, Domain-Specific Applications such as surgical or household robotics, Representation Learning for continuous embeddings of graph elements, and Benchmarking and Datasets that provide standardized evaluation.

A particularly active line of work lies at the intersection of LLM-based planning and scene graph reasoning, where systems like SayPlan[9] and EmbodiedRAG[27] demonstrate how pretrained language models can be grounded in structured spatial knowledge to produce more interpretable and adaptable plans. MomaGraph[0] sits squarely within this LLM-based task planning cluster, emphasizing how multi-modal scene graphs can inform language-driven decision-making for embodied agents. Compared to Hierarchical Action Generation[28], which focuses on decomposing actions into sub-goals, MomaGraph[0] places greater emphasis on the interplay between visual scene understanding and linguistic task specifications. Meanwhile, Context Matters[30] highlights the importance of situational context in grounding, a theme that complements MomaGraph[0]'s approach to integrating rich relational information. Open questions remain around scalability to large, dynamic environments, the trade-off between symbolic and continuous representations, and how to maintain graph consistency under partial observability and real-time constraints.

Related Works in Same Category

The following **4 sibling papers** share the same taxonomy leaf node with the original paper:

1. SayPlan: Grounding Large Language Models using 3D Scene Graphs for Scalable Task Planning

Authors: Rana, Krishan, Krishan Rana, Haviland, Jesse, et al. (20 authors total) | **Year/Venue**: 2023 • Conference on Robot Learning | **URL**: View paper

Abstract

Large language models (LLMs) have demonstrated impressive results in developing generalist planning agents for diverse tasks. However, grounding these plans in expansive, multi-floor, and multi-room environments presents a significant challenge for robotics. We introduce SayPlan, a scalable approach to LLM-based, large-scale task planning for robotics using 3D scene graph (3DSG) representations. To ensure the scalability of our approach, we: (1) exploit the hierarchical nature of 3DSGs to allow ...

Relationship Analysis

Both papers belong to the LLM-Based Task Planning with Scene Graphs category, leveraging large language models grounded in structured scene representations for embodied task planning. They overlap in using scene graphs as intermediate representations to improve LLM-based planning accuracy and robustness in household environments. However, MomaGraph introduces a unified spatial-functional scene graph with part-level nodes and state-aware dynamic updates, trained via reinforcement learning, while SayPlan focuses on hierarchical semantic search over 3D scene graphs with iterative replanning using classical path planners and scene graph simulators for scalability across multi-floor environments.

---

2. EmbodiedRAG: Dynamic 3D Scene Graph Retrieval for Efficient and Scalable Robot Task Planning

Authors: Meghan Booker, Grayson Byrd, Bethany Kemp, Aurora Schmidt, Booker, et al. (9 authors total) | **Year/Venue**: 2024 • arXiv.org | **URL**: View paper

Abstract

Recent advances in Large Language Models (LLMs) have helped facilitate exciting progress for robotic planning in real, open-world environments. 3D scene graphs (3DSGs) offer a promising environment representation for grounding such LLM-based planners as they are compact and semantically rich. However, as the robot's environment scales (e.g., number of entities tracked) and the complexity of scene graph information increases (e.g., maintaining more attributes), providing the 3DSG as-is to an LLM-...

Relationship Analysis

Both papers belong to the LLM-Based Task Planning with Scene Graphs category, using scene graphs as structured representations to ground large language models for embodied task planning. While MomaGraph focuses on generating unified spatial-functional scene graphs with part-level nodes and state-awareness through a trained VLM (MomaGraph-R1), EmbodiedRAG addresses the scalability challenge by retrieving task-relevant subgraphs from existing 3D scene graphs to reduce LLM input token counts. The key difference is that MomaGraph emphasizes scene graph construction and representation design with RL-trained models, whereas EmbodiedRAG assumes pre-existing scene graphs and focuses on efficient retrieval mechanisms for LLM planning.

---

3. Hierarchical Generation of Action Sequence for Service Rots Based on Scene Graph via Large Language Models

Authors: Yu Gu, Guohui Tian, Zhengsong Jiang | **Year/Venue**: 2024 | **URL**: View paper

Abstract

To enhance the task execution capability of home service robots and address the issues of uncontrolled output quality of LLMs and the inability to update information autonomously, we propose the Hierarchical Action Sequence Generation (HGAS) framework for home service robots based on scene graphs and Large Language Models (LLMs). The HGAS framework, which is an innovative approach to action sequence generation by combining the LLMs and the graph editing algorithm, is uniquely designed to process...

Relationship Analysis

Both papers belong to the LLM-Based Task Planning with Scene Graphs category, using large language models grounded in scene graph representations for embodied task planning. They overlap in leveraging scene graphs as structured intermediate representations to improve LLM-based planning for household robots. However, MomaGraph focuses on unified spatial-functional scene graphs with part-level nodes and reinforcement learning optimization, while HGAS emphasizes hierarchical action generation with dynamic graph editing and error correction mechanisms for real-time replanning when actions fail.

---

4. Context Matters! Relaxing Goals with LLMs for Feasible 3D Scene Planning

Authors: Brienza, Michele, Emanuele Musumeci, Michele Brienza, F. Argenziano, et al. (13 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Embodied agents need to plan and act reliably in real and complex 3D environments. Classical planning (e.g., PDDL) offers structure and guarantees, but in practice it fails under noisy perception and incorrect predicate grounding. On the other hand, Large Language Models (LLMs)-based planners leverage commonsense reasoning, yet frequently propose actions that are unfeasible or unsafe. Following recent works that combine the two approaches, we introduce ContextMatters, a framework that fuses LLMs...

Relationship Analysis

Both papers belong to the LLM-Based Task Planning with Scene Graphs category, leveraging large language models grounded in structured scene representations for embodied task planning. They overlap in using scene graphs as intermediate representations to improve planning reliability and in combining LLM reasoning with symbolic planning methods. However, the original paper (MomaGraph) focuses on unified spatial-functional scene graph generation with part-level nodes and state-aware dynamic updates, training a specialized VLM with reinforcement learning, while the candidate paper (Context Matters) emphasizes hierarchical goal relaxation mechanisms that adapt infeasible tasks to contextually achievable objectives through iterative PDDL-based refinement without modifying the underlying scene graph representation.

Contributions Analysis

Overall novelty summary. The paper introduces MomaGraph, a unified scene representation integrating spatial-functional relationships and part-level interactive elements for mobile manipulators. It resides in the 'LLM-Based Task Planning with Scene Graphs' leaf, which contains five papers including the original work. This leaf sits within the broader 'Task Planning and Reasoning with Scene Graphs' branch, indicating a moderately populated research direction. The taxonomy shows that while scene graph construction and navigation have multiple specialized subcategories, task planning with LLMs represents a focused but active area where structured spatial knowledge meets language-driven reasoning.

The taxonomy reveals neighboring work in 'Schema-Guided and Multi-Agent Reasoning' (two papers) and 'Classical and Symbolic Planning with Scene Graphs' (two papers), suggesting that task planning approaches vary in their reliance on foundation models versus symbolic methods. The 'Functional and Interactive Scene Graphs' leaf (four papers) in the construction branch addresses similar concerns about affordances and part-level modeling, though it focuses on representation rather than planning. The 'Dynamic and Temporal Scene Graph Updating' leaf (three papers) tackles temporal consistency, a challenge MomaGraph addresses through its emphasis on object states and updates, bridging construction and planning concerns.

Across three contributions, the analysis examined thirty candidates total, with zero refutable pairs identified. The unified representation contribution examined ten candidates with none providing clear overlap; the vision-language model contribution similarly found no refutations among ten candidates; and the dataset-benchmark contribution encountered no prior work among its ten examined papers. This limited search scope—thirty semantically similar papers from a fifty-paper taxonomy—suggests the analysis captures closely related work but may not reflect the full breadth of scene graph research. The absence of refutations indicates that among these top matches, no single prior work directly anticipates MomaGraph's specific combination of spatial-functional integration, part-level modeling, and task-driven evaluation.

Given the search examined roughly sixty percent of the taxonomy's papers through semantic similarity, the findings suggest MomaGraph occupies a relatively distinct position within its immediate neighborhood. The lack of refutations across all contributions, combined with the paper's placement in a moderately populated leaf, implies the work synthesizes ideas from multiple branches—construction, planning, and benchmarking—in a novel configuration. However, the limited scope means potential overlaps in the broader literature, particularly in functional scene graphs or memory-augmented planning, remain unexplored by this analysis.

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

Contribution 1: MomaGraph: Unified Scene Graph Representation

Description: MomaGraph is a novel scene representation that unifies spatial and functional relationships while introducing part-level interactive nodes. It provides a compact, adaptive, and task-relevant structured representation for embodied agents, addressing limitations of prior work that separated spatial and functional relations or treated scenes as static snapshots.

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. FunGraph: Functionality Aware 3D Scene Graphs for Language-Prompted Scene Interaction

URL: View paper

Brief Assessment

FunGraph[56] focuses on affordance-aware 3D scene graphs for object manipulation with part-level functional elements in robotics contexts, while MomaGraph addresses unified spatial-functional relationships for embodied task planning with state-aware dynamic updates. The technical approaches and application domains differ substantially.

---

2. Part-level scene reconstruction affords robot interaction

URL: View paper

Brief Assessment

Part Level Reconstruction[58] focuses on part-level geometric reconstruction of objects using primitive shapes for robot manipulation, not on unified scene graph representations that jointly model spatial-functional relationships for task planning.

---

3. 3D scene graphs in robotics: A unified representation bridging geometry, semantics, and action

URL: View paper

Brief Assessment

Scene Graphs Robotics[15] focuses on general 3D scene graph representations in robotics bridging geometry, semantics, and action, but does not specifically address the unified modeling of spatial and functional relationships with part-level interactive nodes as proposed in MomaGraph.

---

4. Visual knowledge graph for human action reasoning in videos

URL: View paper

Brief Assessment

Visual Knowledge Graph[55] focuses on action recognition in videos through body part movements and interactive objects, not on unified spatial-functional scene representations for embodied agents in 3D indoor environments.

---

5. SceneHGN: Hierarchical Graph Networks for 3D Indoor Scene Generation With Fine-Grained Geometry

URL: View paper

Brief Assessment

SceneHGN[57] focuses on hierarchical generation of 3D indoor scenes with fine-grained mesh geometry, not on unified scene graph representations for embodied task planning with state-aware dynamic updates.

---

6. Dynamic open-vocabulary 3d scene graphs for long-term language-guided mobile manipulation

URL: View paper

Brief Assessment

Dynamic Open Vocabulary Graphs[51] focuses on dynamic scene graph updates for mobile manipulation in changing environments, not on the unified spatial-functional representation with part-level nodes that MomaGraph introduces. The candidate emphasizes real-time adaptation to environmental changes rather than the joint modeling of spatial and functional relationships at the part level.

---

7. Scene Graph Generation with Role-Playing Large Language Models

URL: View paper

Brief Assessment

Scene-Specific Description based Scene Graph Generation (Role Playing Scene Graphs[52]) focuses on open-vocabulary scene graph generation using adaptive text classifiers and LLM-generated descriptions, not on unified spatial-functional representations with part-level interactive nodes for embodied task planning.

---

8. Fast Contextual Scene Graph Generation with Unbiased Context Augmentation

URL: View paper

Brief Assessment

Fast Contextual Generation[54] focuses on scene graph generation for visual relationship prediction in static images, addressing long-tail bias and inference speed. It does not address embodied agents, spatial-functional relationship unification, part-level interactive nodes, or dynamic state-aware updates that characterize MomaGraph.

---

9. Commonsense scene graph-based target localization for object search

URL: View paper

Brief Assessment

Commonsense Target Localization[5] focuses on object search tasks using scene graphs for target localization in household environments, not on proposing a general unified scene graph representation framework for embodied task planning.

---

10. Hydra: A Real-time Spatial Perception System for 3D Scene Graph Construction and Optimization

URL: View paper

Brief Assessment

Hydra[53] focuses on real-time 3D scene graph construction from sensor data with topological mapping and loop closure, not on unifying spatial-functional relationships with part-level interactive nodes for task planning.

Contribution 2: MomaGraph-R1: Vision-Language Model with Reinforcement Learning

Description: MomaGraph-R1 is a 7B vision-language model trained using the DAPO reinforcement learning algorithm with a graph-alignment reward function. It generates task-oriented scene graphs and serves as a zero-shot task planner within a Graph-then-Plan framework, improving reasoning effectiveness and interpretability.

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. Spatial-SSRL: Enhancing Spatial Understanding via Self-Supervised Reinforcement Learning

URL: View paper

Brief Assessment

Spatial-SSRL[70] focuses on self-supervised reinforcement learning for general spatial understanding tasks (depth ordering, relative position) using intrinsic image consistency signals, whereas MomaGraph-R1 targets task-oriented scene graph generation for embodied task planning with graph-alignment rewards in household manipulation scenarios.

---

2. Brain-Inspired Planning for Better Generalization in Reinforcement Learning

URL: View paper

Brief Assessment

Brain Inspired Planning[73] focuses on enhancing RL agents' zero-shot systematic generalization through brain-inspired reasoning mechanisms like spatial abstraction and task decomposition, not on vision-language models for scene graph generation and spatial-functional reasoning in embodied task planning.

---

3. Prompt Informed Reinforcement Learning for Visual Coverage Path Planning

URL: View paper

Brief Assessment

Prompt Informed Coverage[67] focuses on UAV visual coverage path planning using LLM-guided reward shaping for PPO, not on vision-language models for spatial-functional reasoning or scene graph generation for embodied task planning.

---

4. Atari-GPT: Benchmarking Multimodal Large Language Models as Low-Level Policies in Atari Games

URL: View paper

Brief Assessment

Atari GPT[68] focuses on evaluating multimodal LLMs as low-level controllers in Atari games, not on training vision-language models with RL for spatial-functional reasoning and scene graph generation in embodied task planning.

---

5. Multimodal Visual Transformer for Sim2real Transfer in Visual Reinforcement Learning

URL: View paper

Brief Assessment

Multimodal Visual Transformer[71] focuses on RGB-depth fusion for visual reinforcement learning in robotic manipulation with sim2real transfer, not on vision-language models for spatial-functional reasoning or scene graph generation for task planning.

---

6. UAV-VL-R1: Generalizing Vision-Language Models via Supervised Fine-Tuning and Multi-Stage GRPO for UAV Visual Reasoning

URL: View paper

Brief Assessment

UAV VL R1[65] focuses on aerial imagery reasoning for UAV applications with GRPO-based RL, while MomaGraph-R1 targets embodied task planning in household environments using DAPO RL with graph-alignment rewards. The domains, tasks, and RL formulations differ fundamentally.

---

7. MomaGraph: State-Aware Unified Scene Graphs with Vision-Language Model for Embodied Task Planning

URL: View paper

Brief Assessment

MomaGraph State Aware[72] is the same paper as the original submission. The candidate paper title explicitly references the original work, indicating this is the same contribution rather than prior work that could refute novelty claims.

---

8. Ariadne: A Controllable Framework for Probing and Extending VLM Reasoning Boundaries

URL: View paper

Brief Assessment

Ariadne[69] focuses on spatial reasoning in synthetic mazes using RL for curriculum learning, not on scene graph generation or embodied task planning in household environments.

---

9. Sense, Imagine, Act: Multimodal Perception Improves Model-Based Reinforcement Learning for Head-to-Head Autonomous Racing

URL: View paper

Brief Assessment

Sense Imagine Act[66] focuses on model-based reinforcement learning for autonomous racing using multimodal perception (lidar and camera fusion), not vision-language models for spatial-functional reasoning or scene graph generation for embodied task planning.

---

10. SeeNav-Agent: Enhancing Vision-Language Navigation with Visual Prompt and Step-Level Policy Optimization

URL: View paper

Brief Assessment

SeeNav Agent[74] focuses on vision-language navigation tasks with step-level policy optimization for navigation success, not on scene graph generation or spatial-functional reasoning for embodied task planning.

Contribution 3: MomaGraph-Scenes Dataset and MomaGraph-Bench Evaluation Suite

Description: MomaGraph-Scenes is the first dataset jointly modeling spatial and functional relationships with part-level annotations, encompassing multi-view observations and task-aligned scene graphs. MomaGraph-Bench is a comprehensive benchmark evaluating six reasoning capabilities from high-level planning to fine-grained scene understanding.

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. Reasoning with scene graphs for robot planning under partial observability

URL: View paper

Brief Assessment

Reasoning Partial Observability[20] focuses on robot planning under partial observability using scene graphs for target search tasks, not on creating task-driven scene graph datasets with multi-view observations and comprehensive reasoning benchmarks for embodied planning.

---

2. Taskography: Evaluating robot task planning over large 3d scene graphs

URL: View paper

Brief Assessment

Taskography[64] focuses on symbolic task planning over large 3D scene graphs in building-scale environments, while the original paper presents task-driven scene graphs with spatial-functional relationships and part-level annotations for household manipulation tasks. The candidate addresses different planning scales and does not challenge the novelty of joint spatial-functional modeling with part-level interactive elements.

---

3. Optimal scene graph planning with large language model guidance

URL: View paper

Brief Assessment

Optimal Scene Graph Planning[61] focuses on translating natural language to LTL formulas for task planning in known scene graphs, not on constructing task-driven scene graph datasets with part-level annotations or comprehensive reasoning benchmarks.

---

4. Exploring 3D Reasoning-Driven Planning: From Implicit Human Intentions to Route-Aware Activity Planning

URL: View paper

Brief Assessment

Reasoning Driven Planning[63] focuses on reasoning implicit user intentions and route planning for robot moves, while MomaGraph-Scenes emphasizes joint spatial-functional relationships with part-level annotations for task-driven scene graphs. The datasets serve different purposes and evaluation goals.

---

5. EmbodiedVSR: Dynamic Scene Graph-Guided Chain-of-Thought Reasoning for Visual Spatial Tasks

URL: View paper

Brief Assessment

EmbodiedVSR[18] introduces the espatial-benchmark for spatial reasoning evaluation, which differs from MomaGraph-Scenes' focus on jointly modeling spatial and functional relationships with part-level annotations and task-aligned scene graphs for embodied planning.

---

6. Embodied agent interface: Benchmarking llms for embodied decision making

URL: View paper

Brief Assessment

Embodied Agent Interface[60] focuses on benchmarking LLMs for embodied decision-making through standardized task interfaces and evaluation metrics, not on creating task-driven scene graph datasets with spatial-functional relationships and part-level annotations.

---

7. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning

URL: View paper

Brief Assessment

Sayplan Grounding[59] focuses on using pre-constructed 3D scene graphs for robot task planning in large-scale environments, not on creating datasets with part-level annotations or comprehensive benchmarks for evaluating reasoning capabilities. The candidate addresses scalability of planning with existing scene graphs, while the original contribution is about dataset construction and evaluation methodology.

---

8. Grapheqa: Using 3d semantic scene graphs for real-time embodied question answering

URL: View paper

Brief Assessment

Grapheqa[3] focuses on embodied question answering using real-time 3D scene graphs for navigation and visual grounding, not on creating task-driven scene graph datasets with part-level annotations or comprehensive reasoning benchmarks spanning high-level planning to fine-grained understanding.

---

9. Schema-Guided Scene-Graph Reasoning based on Multi-Agent Large Language Model System

URL: View paper

Brief Assessment

Schema Guided Reasoning[16] focuses on multi-agent LLM reasoning over existing scene graphs for Q&A and planning tasks, not on creating datasets with joint spatial-functional annotations or comprehensive benchmarks evaluating six reasoning capabilities from planning to scene understanding.

---

10. Verigraph: Scene graphs for execution verifiable robot planning

URL: View paper

Brief Assessment

Verigraph[62] focuses on execution verification for robot planning using scene graphs in manipulation tasks, not on creating task-driven datasets with part-level annotations or comprehensive reasoning benchmarks spanning multiple capabilities from planning to scene understanding.

Appendix: Text Similarity Detection

Textual similarity detection checked 34 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. MomaGraph: State-Aware Unified Scene Graphs with Vision-Language Model for Embodied Task Planning

Detected in: Contribution: contribution_2

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] MomaGraph: State-Aware Unified Scene Graphs with Vision-Language Models for Embodied Task Planning View paper
- [1] Open-vocabulary functional 3d scene graphs for real-world indoor spaces View paper
- [2] Roboexp: Action-conditioned scene graph via interactive exploration for robotic manipulation View paper
- [3] Grapheqa: Using 3d semantic scene graphs for real-time embodied question answering View paper
- [4] VLM-MSGraph: Vision Language Model-enabled Multi-hierarchical Scene Graph for robotic assembly View paper
- [5] Commonsense scene graph-based target localization for object search View paper
- [6] Mesatask: Towards task-driven tabletop scene generation via 3d spatial reasoning View paper
- [7] Learning Bird's Eye View scene graph and knowledge-inspired policy for embodied visual navigation View paper
- [8] In defense of scene graph generation for human-robot open-ended interaction in service robotics View paper
- [9] SayPlan: Grounding Large Language Models using 3D Scene Graphs for Scalable Task Planning View paper
- [10] Agentworld: An interactive simulation platform for scene construction and mobile robotic manipulation View paper
- [11] TACS-Graphs: Traversability-Aware Consistent Scene Graphs for Ground Robot Indoor Localization and Mapping View paper
- [12] Scene Graph Contrastive Learning for Embodied Navigation View paper
- [13] Long-term robot manipulation task planning with scene graph and semantic knowledge View paper
- [14] Dynamic Scene Generation for Embodied Navigation Benchmark View paper
- [15] 3D scene graphs in robotics: A unified representation bridging geometry, semantics, and action View paper
- [16] Schema-Guided Scene-Graph Reasoning based on Multi-Agent Large Language Model System View paper
- [17] Scene Augmentation Methods for Interactive Embodied AI Tasks View paper
- [18] EmbodiedVSR: Dynamic Scene Graph-Guided Chain-of-Thought Reasoning for Visual Spatial Tasks View paper
- [19] SG-Nav: Online 3D Scene Graph Prompting for LLM-based Zero-shot Object Navigation View paper
- [20] Reasoning with scene graphs for robot planning under partial observability View paper
- [21] Generation of skill-specific maps from graph world models for robotic systems View paper
- [22] Scenethesis: A language and vision agentic framework for 3d scene generation View paper
- [23] Dynamic Interactive Relation Capturing via Scene Graph Learning for Robotic Surgical Report Generation View paper
- [24] Hi-Dyna Graph: Hierarchical Dynamic Scene Graph for Robotic Autonomy in Human-Centric Environments View paper
- [25] Scene-Driven Multimodal Knowledge Graph Construction for Embodied AI View paper
- [26] Information-Theoretic Graph Fusion with Vision-Language-Action Model for Policy Reasoning and Dual Robotic Control View paper
- [27] EmbodiedRAG: Dynamic 3D Scene Graph Retrieval for Efficient and Scalable Robot Task Planning View paper

- [28] Hierarchical Generation of Action Sequence for Service Rots Based on Scene Graph via Large Language Models View paper
- [29] Agentic 3D Scene Generation with Spatially Contextualized VLMs View paper
- [30] Context Matters! Relaxing Goals with LLMs for Feasible 3D Scene Planning View paper
- [31] General Scene Adaptation for Vision-and-Language Navigation View paper
- [32] Time is on my sight: scene graph filtering for dynamic environment perception in an LLM-driven robot View paper
- [33] GraphAD: Interaction Scene Graph for End-to-end Autonomous Driving View paper
- [34] GeoNav: Empowering MLLMs with Explicit Geospatial Reasoning Abilities for Language-Goal Aerial Navigation View paper
- [35] SituationalLLM: Proactive Language Models with Scene Awareness for Dynamic, Contextual Task Guidance View paper
- [36] Task-Driven Graph Attention for Hierarchical Relational Object Navigation View paper
- [37] Lost & Found: Tracking Changes From Egocentric Observations in 3D Dynamic Scene Graphs View paper
- [38] 3D dynamic scene graphs: Actionable spatial perception with places, objects, and humans View paper
- [39] Belief Scene Graphs: Expanding Partial Scenes with Objects through Computation of Expectation View paper
- [40] Orionnav: Online planning for robot autonomy with context-aware llm and open-vocabulary semantic scene graphs View paper
- [41] SG-Bot: Object Rearrangement via Coarse-to-Fine Robotic Imagination on Scene Graphs View paper
- [42] SOON: Scenario Oriented Object Navigation with Graph-based Exploration View paper
- [43] Graph-based Task-specific Prediction Models for Interactions between Deformable and Rigid Objects View paper
- [44] Continuous Scene Representations for Embodied AI View paper
- [45] KARMA: Augmenting Embodied AI Agents with Long-and-Short Term Memory Systems View paper
- [46] Open scene graphs for open-world object-goal navigation View paper
- [47] VISAGE: Video Synthesis using Action Graphs for Surgery View paper
- [48] Mapping High-level Semantic Regions in Indoor Environments without Object Recognition View paper
- [49] Multi-modal scene graph inspired policy for visual navigation View paper
- [50] Knowledge-driven Scene Priors for Semantic Audio-Visual Embodied Navigation View paper
- [51] Dynamic open-vocabulary 3d scene graphs for long-term language-guided mobile manipulation View paper
- [52] Scene Graph Generation with Role-Playing Large Language Models View paper
- [53] Hydra: A Real-time Spatial Perception System for 3D Scene Graph Construction and Optimization View paper
- [54] Fast Contextual Scene Graph Generation with Unbiased Context Augmentation View paper
- [55] Visual knowledge graph for human action reasoning in videos View paper
- [56] FunGraph: Functionality Aware 3D Scene Graphs for Language-Prompted Scene Interaction View paper
- [57] SceneHGN: Hierarchical Graph Networks for 3D Indoor Scene Generation With Fine-Grained Geometry View paper
- [58] Part-level scene reconstruction affords robot interaction View paper
- [59] Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning View paper
- [60] Embodied agent interface: Benchmarking llms for embodied decision making View paper
- [61] Optimal scene graph planning with large language model guidance View paper
- [62] Verigraph: Scene graphs for execution verifiable robot planning View paper
- [63] Exploring 3D Reasoning-Driven Planning: From Implicit Human Intentions to Route-Aware Activity Planning View paper
- [64] Taskography: Evaluating robot task planning over large 3d scene graphs View paper
- [65] UAV-VL-R1: Generalizing Vision-Language Models via Supervised Fine-Tuning and Multi-Stage GRPO for UAV Visual Reasoning View paper
- [66] Sense, Imagine, Act: Multimodal Perception Improves Model-Based Reinforcement Learning for Head-to-Head Autonomous Racing View paper
- [67] Prompt Informed Reinforcement Learning for Visual Coverage Path Planning View paper
- [68] Atari-GPT: Benchmarking Multimodal Large Language Models as Low-Level Policies in Atari Games View paper
- [69] Ariadne: A Controllable Framework for Probing and Extending VLM Reasoning Boundaries View paper
- [70] Spatial-SSRL: Enhancing Spatial Understanding via Self-Supervised Reinforcement Learning View paper
- [71] Multimodal Visual Transformer for Sim2real Transfer in Visual Reinforcement Learning View paper
- [72] MomaGraph: State-Aware Unified Scene Graphs with Vision-Language Model for Embodied Task Planning View paper
- [73] Brain-Inspired Planning for Better Generalization in Reinforcement Learning View paper
- [74] SeeNav-Agent: Enhancing Vision-Language Navigation with Visual Prompt and Step-Level Policy Optimization View paper