Project Group/Seminar/Lab
WT/ST [Year]
Title

Student Name
Matriculation No: 0000000

Supervisor: Supervisor Name

2026-01-18

Abstract

To fully leverage mobile manipulation robots, agents must autonomously execute long-horizon tasks in large, unexplored environments [1]. This paper addresses the challenge of interactive object search, where robots must navigate, explore, and manipulate objects (e.g., opening doors and drawers) to find targets in partially observable settings [2]. We propose MoMa-LLM, a novel approach that grounds Large Language Models (LLMs) within structured representations derived from open-vocabulary scene graphs, which are dynamically updated during exploration [3]. By tightly interleaving these representations with an object-centric action space, the system achieves zero-shot, open-vocabulary reasoning [4]. To rigorously benchmark performance, we introduce a novel evaluation paradigm utilizing full efficiency curves and the Area Under the Curve for Efficiency (AUC-E) metric [5]. Extensive experiments in simulation and the real world demonstrate that MoMa-LLM substantially improves search efficiency compared to state-of-the-art baselines like HIMOS and ESC-Interactive [6].

1 Introduction

Interactive embodied AI tasks in large, human-centered environments require robots to reason over long horizons and interact with a multitude of objects [7]. In many real-world scenarios, these environments are a priori unknown or continuously rearranged, making autonomous operation significantly more difficult [8]. Specifically, the task of interactive object search presents unique challenges because objects are not always openly visible; they may be stored inside receptacles like cabinets or drawers, or located behind closed doors [9]. Consequently, an agent cannot rely on directional reasoning alone but must actively manipulate the environment—opening doors to navigate and searching through containers—to succeed [10].

Recent advancements have shown the potential of Large Language Models (LLMs) for generating high-level robotic plans [11]. However, existing methods such as SayCan or SayPlan are insufficient for interactive search in unexplored settings for several reasons:

- **Assumption of Full Observability:** Many prior works focus on fully observed environments, such as tabletop manipulation or pre-explored scenes [12].
- **Scalability and Hallucination:** In large, partially observable scenes with numerous objects, simply providing an LLM with raw observations or lists of objects increases the risk of generating impractical sequences or hallucinations [13]. Simple prompting strategies or raw JSON inputs of full scene graphs have proven insufficient for complex reasoning in these contexts [14].
- **Limited Scope:** Methods often restrict tasks to single rooms or rely on non-interactive navigation, failing to address the complexities of multi-room exploration and physical interaction [15].

1.1 Proposed Solution

To address these limitations, the authors propose MoMa-LLM, a method that grounds LLMs in dynamically built scene graphs [16]. This approach utilizes a scene understanding module that constructs open-vocabulary scene graphs from dense maps and Voronoi graphs as the robot explores [17]. By extracting structured and compact textual representations from these dynamic graphs, the system enables pre-trained LLMs to plan efficiently in partially observable environments [18]. This allows the robot to perform zero-shot, open-vocabulary reasoning, extending readily to a spectrum of complex mobile manipulation tasks [19].

2 Related Work

2.1 3D Scene Graphs

- **Hydra:** Represents the state-of-the-art in real-time 3D Scene Graph construction, abstracting geometry into topology [1].
- **Other:** Approaches like ConceptGraphs and VoroNav investigate zero-shot perception inputs for task planning [2].

2.2 LLMs for Planning

- **SayPlan:** Focuses on identifying subgraphs within large, known scene graphs for planning [5].
- **Other:** LLM-Planner and other methods often restrict tasks to fully observed environments or single rooms [6].

2.3 Object Search

- **ESC:** A baseline for exploration with soft commonsense constraints, scoring frontiers based on object co-occurrences [9].
- **Other:** Classical frontier exploration and reinforcement learning methods like HIMOS (which uses hierarchical RL) [10].

3 Paper Summary

3.1 Problem Statement

The authors address the challenge of embodied reasoning where a robotic agent must locate specific objects within large, unexplored environments.

POMDP formulation The problem is formalized as a Partially Observable Markov Decision Process (POMDP) tuple $\mathcal{M} = (S, A, O, T, P, r)$:

- **S (States):** The complete state of the world, including robot pose, map, and object states (e.g., drawer open/closed) [3].
- **A (Actions):** Hybrid action space with high-level primitives (navigate, open) and low-level controllers [5].
- **O (Observations):** RGB-D images, robot pose, and semantic segmentations received by the agent [6].
- **T (Transition):** $T(s'|s, a)$ representing the probability of the state evolving from s to s' given action a [7].
- **P (Observation model):** $P(o|s)$ representing the likelihood of receiving observation o in state s [8].

- **r (Reward):** Optimizing for efficient search (finding the object with minimal cost) [9].

3.2 Approach: MoMa-LLM

MoMa-LLM grounds Large Language Models (LLMs) in dynamically built scene graphs to enable zero-shot object search.

3.2.1 Hierarchical 3D Scene Graph

The system builds a layered representation of the environment.

Voronoi Graph Construction

- **Construction:** Created from the Generalized Voronoi Diagram (GVD) of the inflated occupancy map, representing points equidistant to obstacles [7].
- **Utility:** Provides a safe navigational backbone that maximizes clearance from obstacles, critical for mobile manipulators [10].

Room Classification

- **Prompting:** The LLM is provided with a list of object categories detected within a room cluster (e.g., "bed, lamp") [19].

3.2.2 High-Level Action Space

The agent operates using the following discrete high-level actions:

- `Maps(room, object)`: Navigate to a specific object or location.
- `go_to_and_open(room, object)`: Navigate to and interact with a container (e.g., fridge).
- `close(room, object)`: Close an opened container.
- `explore(room)`: Visit unexplored frontiers associated with a specific room.
- `done()`: Declare the task explicitly successful.

3.2.3 Grounded High-Level Planning

The system converts the dynamic graph into a structured text prompt for the LLM.

Scene Structure Encoding

- **Serialization:** The graph is serialized into a hierarchical list (Room -> Objects).
- **Abstraction:** Distances are binned into natural language adjectives ("near", "far") and object states are explicit ("closed fridge") to facilitate reasoning [1].

Partial Observability

- **Frontiers:** Unknown space is explicitly represented as "Unexplored Area" (local) or "Leading Out" (global) nodes, allowing the LLM to reason about exploration [6].
- **Replanning:** The prompt includes an "Analysis" and "Reasoning" step (Chain-of-Thought) before generating the next action command, with history dynamically realigned to the current map [14].

3.3 Experiments

3.3.1 Experimental Setup

- **Simulator:** iGibson (based on real-world scans) [1].
- **Robot (Sim):** Fetch mobile manipulator [2].
- **Robot (Real):** Toyota HSR [3].

3.3.2 Evaluation Metrics

- **Success Rate (SR):** Percentage of episodes where the target object is successfully found and `done()` is called.
- **SPL:** Success weighted by Path Length, measuring navigation efficiency.
- **AUC-E (Area Under Efficiency Curve):** A novel metric introduced to evaluate search efficiency across varying time budgets, providing a scalar value that rewards fast and reliable agents, unlike binary cutoffs [8].

3.3.3 Results

Simulation

- **MoMa-LLM vs Baselines:** MoMa-LLM achieved the highest performance (SR 97.7%, AUC-E 87.2%), significantly outperforming unstructured LLM baselines and RL-based methods like HIMOS (SPL 48.5) [13].
- **Failure Modes:** Primary failures were due to perception limitations or "infeasible actions" generated by the LLM when context was overwhelmed, though structured grounding reduced this significantly compared to baselines [15].

Real-world Transfer

- **Transfer Capabilities:** The high-level reasoning transferred effectively to the real world (80% success rate), showing robustness to domain shifts in perception [18].
- **Latency and Constraints:** Main constraints involved the computational latency of LLM queries and the reliance on specific detectors (like AR markers) to handle real-world perception noise [19].

4 Discussion

Strengths

- **Open-vocabulary reasoning:** Capable of handling novel room and object categories without retraining [1].
- **Structured reasoning:** Using scene graphs as a bridge allows the LLM to reason effectively about geometry and topology [2].
- **Robustness:** Demonstrated resilience to segmentation errors and real-world domain shifts [4].

Limitations

- **Perception dependency:** Heavily relies on ground-truth or high-quality semantics; limited by "garbage-in/garbage-out" [5].
- **Latency:** LLM queries dominate the runtime, and full graph re-computation scales poorly [9].
- **Cost blindness:** High-level planning decouples reasoning from low-level execution costs [12].

Reproducibility

- **Code availability:** The authors reference a project page, but specific release details for the full pipeline are limited [14].
- **Missing details:** Real-world failure analysis is constrained by limited feedback modalities [16].

5 Conclusion

MoMa-LLM establishes a robust standard for semantic grounding in exploration by treating the Scene Graph as a dynamic, structured prompt. It effectively operationalizes the "world knowledge" of LLMs for robotics, proving that structured data organization is key to bridging the grounding gap. While issues of specific perception requirements and latency remain, the architectural blueprint of dynamic, language-grounded graphs marks a definitive shift towards hybrid neuro-symbolic architectures in embodied AI [1]. Future work lies in cost-aware grounding and integrating VLM-based room classification [13].

References

[1] e. a. Honerkamp, "Language-grounded dynamic scene graphs for interactive object search with mobile manipulation," arXiv preprint arXiv:2403.08605, 2024. [Online]. Available: https://arxiv.org/abs/2403.08605

[2] e. a. Schmalstieg, "Learning long-horizon robot exploration strategies for multi-object search in continuous action spaces," *Proc. of ICRA*, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Learning-Long-Horizon-Robot-Exploration-Strategies-Schmalstieg-Honerkamp/915234d691be2d9efdbdcc178fbff28a381c0863

[3] ——, "Learning hierarchical interactive multi-object search for mobile manipulation," *Proc. of IROS*, 2023. [Online]. Available: https://www.researchgate.net/publication/372313553_Learning_Hierarchical_Interactive_Multi-Object_Search_for_Mobile_Manipulation

[4] e. a. Mohammadi, "More: Mobile manipulation rearrangement through grounded language reasoning," Preprint, 2025. [Online]. Available: https://tisl.cs.toronto.edu/publication/MORE__Mobile_Manipulation_Rearrangement_Through_Grounded_Language_Reasoning/MORE__Mobile_Manipulation_Rearrangement_Through_Grounded_Language_Reasoning.pdf

[5] e. a. Honerkamp, "Moma-llm project page," 2024. [Online]. Available: https://moma-llm.cs.uni-freiburg.de/

[6] e. a. Yang, "Esc: Exploration with soft commonsense constraints for zero-shot object navigation," in *ICRA*, 2023. [Online]. Available: https://openreview.net/pdf?id=GydFM0ZEXY

[7] e. a. Honerkamp, "Language-grounded dynamic scene graphs for interactive object search with mobile manipulation," arXiv preprint arXiv:2403.08605, 2024. [Online]. Available: https://arxiv.org/abs/2403.08605

[8] ——, "Language-grounded dynamic scene graphs for interactive object search with mobile manipulation," *arXiv preprint*, 2024. [Online]. Available: https://openreview.net/pdf/f0ccb8e5511b7ca0d776b046872e4ed72e6e3233.pdf

[9] ——, "Language-grounded dynamic scene graphs for interactive object search with mobile manipulation," in *RSS Workshop on Semantic Robotics*, 2024. [Online]. Available: https://semrob.github.io/docs/rss_semrob2024_cr_paper26.pdf

[10] e. a. Author, "Lookplangraph: Embodied instruction following method with vlm graph augmentation," in *OpenReview*, 2024. [Online]. Available: https://openreview.net/pdf?id=B47cCZfJFa

[11] e. a. Yang, "Esc: Exploration with soft commonsense constraints for zero-shot object navigation (arxiv)," arXiv preprint arXiv:2301.13166, 2023. [Online]. Available: https://arxiv.org/abs/2301.13166

[12] e. a. Honerkamp, "Language-grounded dynamic scene graphs for interactive object search with mobile manipulation," Scribd Document, 2024. [Online]. Available: https://www.scribd.com/document/787749169/2403-08605v4

[13] e. a. Author, "N^2m^2: Learning navigation for arbitrary mobile manipulation motions in unseen and dynamic environments," in *Proc. of IROS*, 2023. [Online]. Available: https://www.semanticscholar.org/paper/5051947cbac4223a68edc8e7e7319b5cdb2dc712

[14] e. a. Honerkamp, "Moma-llm: Scene graphs for mobile object search (emergent mind)," 2024. [Online]. Available: https://www.emergentmind.com/papers/2403.08605

[15] e. a. Yang, "Interleaved llm and motion planning for generalized multi-object collection in large scene graphs," arXiv preprint arXiv:2507.15782, 2025. [Online]. Available: https://arxiv.org/html/2507.15782v1

[16] ——, "Interleaved llm and motion planning for generalized multi-object collection in large scene graphs (abstract)," arXiv preprint arXiv:2507.15782, 2025. [Online]. Available: https://arxiv.org/abs/2507.15782

[17] e. a. Chen, "Explainable saliency: Articulating reasoning with contextual prioritization," in *Proceedings of CVPR*, 2025. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2025/papers/Chen_Explainable_Saliency_Articulating_Reasoning_with_Contextual_Prioritization_CVPR_2025_paper.pdf

[18] Z. Irshad, "Awesome-robotics-3d github repository," 2024. [Online]. Available: https://github.com/zubair-irshad/Awesome-Robotics-3D

[19] e. a. Author, "Open-vocabulary and semantic-aware reasoning for search and retrieval of objects in dynamic and concealed spaces," *Autonomous Robots*, 2025. [Online]. Available: https://autonomousrobots.nl/assets/images/workshops/2025_iros/accepted_papers/paper_8_Open.pdf