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 [1]. 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 [1]. By tightly interleaving these representations with an object-centric action space, the system achieves zero-shot, open-vocabulary reasoning [1]. 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 [1]. 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 [2] and ESC-Interactive [3].

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 [1]. In many real-world scenarios, these environments are a priori unknown or continuously rearranged, making autonomous operation significantly more difficult [1]. 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 [1]. 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 [2].

Recent advancements have shown the potential of Large Language Models (LLMs) for generating high-level robotic plans [4]. 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 [5].
- **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 [1].
- **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 [2].

1.1 Proposed Solution

To address these limitations, the authors propose MoMa-LLM, a method that grounds LLMs in dynamically built scene graphs [1]. This approach utilizes a scene understanding module that constructs open-vocabulary scene graphs from dense maps and Voronoi graphs as the robot explores [1]. By extracting structured and compact textual representations from these dynamic graphs, the system enables pre-trained LLMs to plan efficiently in partially observable environments [1].

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 [6].
- **Other:** Approaches like ConceptGraphs [7] and VoroNav [8] investigate zero-shot perception inputs for task planning.

2.2 LLMs for Planning

- **SayPlan:** Focuses on identifying subgraphs within large, known scene graphs for planning [5].
- **SayCan:** Grounds LLMs in robotic affordances but focuses on fully observed table-top settings [4].

2.3 Object Search

- **ESC:** A baseline for exploration with soft commonsense constraints, scoring frontiers based on object co-occurrences [3].
- **HIMOS:** A hierarchical reinforcement learning approach for interactive search [2].

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 [1].

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 [1].
- **A (Actions):** Hybrid action space with high-level primitives (navigate, open) and low-level controllers [1].
- **O (Observations):** RGB-D images, robot pose, and semantic segmentations [1].
- **T (Transition):** $T(s'|s, a)$ representing state evolution [1].
- **P (Observation model):** $P(o|s)$ representing observation likelihood [1].
- **r (Reward):** Optimizing for efficient search [1].

3.2 Approach: MoMa-LLM

MoMa-LLM grounds Large Language Models in dynamically built scene graphs [1].

3.2.1 Hierarchical 3D Scene Graph

Voronoi Graph Construction

- **Construction:** Created from the Generalized Voronoi Diagram (GVD) of the inflated occupancy map [1].
- **Utility:** Provides a safe navigational backbone that maximizes clearance from obstacles [1].

Room Classification

- **Prompting:** The LLM is provided with a list of object categories detected within a room cluster [1].

3.2.2 High-Level Action Space

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

3.2.3 Grounded High-Level Planning

Scene Structure Encoding

- **Serialization:** The graph is serialized into a hierarchical list (Room -> Objects) [1].
- **Abstraction:** Distances are binned into natural language adjectives [1].

Partial Observability

- **Frontiers:** Unknown space is explicitly represented as "Unexplored Area" nodes [1].
- **Replanning:** The prompt includes Chain-of-Thought reasoning with dynamic history realignment [1].

3.3 Experiments

3.3.1 Experimental Setup

- **Simulator:** iGibson (based on real-world scans) [9].
- **Robot (Sim):** Fetch mobile manipulator [1].
- **Robot (Real):** Toyota HSR [1].

3.3.2 Evaluation Metrics

- **Success Rate (SR):** Percentage of episodes where the target object is found.
- **SPL:** Success weighted by Path Length.
- **AUC-E:** Area Under the Efficiency Curve, a novel metric [1].

3.3.3 Results

Simulation

- **MoMa-LLM vs Baselines:** MoMa-LLM achieved SR 97.7%, SPL 63.6, AUC-E 87.2 [1].
- **Failure Modes:** Primary failures due to perception limitations [1].

Real-world Transfer

- **Transfer Capabilities:** 80% success rate in real-world experiments [1].
- **Latency:** Main constraints were LLM query latency [1].

4 Discussion

Strengths

- **Open-vocabulary reasoning:** Handles novel categories without retraining [1].
- **Structured reasoning:** Scene graphs enable LLM reasoning about geometry [1].
- **Robustness:** Resilient to segmentation errors and domain shifts [1].

Limitations

- **Perception dependency:** Relies on high-quality semantics [1].
- **Latency:** LLM queries dominate runtime [1].
- **Cost blindness:** High-level planning decoupled from execution costs [1].

Reproducibility

- **Code availability:** Project page available [10].
- **Missing details:** Real-world failure analysis limited [1].

5 Conclusion

MoMa-LLM establishes a robust standard for semantic grounding in exploration by treating the Scene Graph as a dynamic, structured prompt [1]. 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 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. Future work lies in cost-aware grounding and integrating VLM-based room classification [1].

References

[1] D. Honerkamp, T. Schmalstieg, T. Welschehold, N. Abdo, and A. Valada, "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] T. Schmalstieg, D. Honerkamp, T. Welschehold, and A. Valada, "Learning hierarchical interactive multi-object search for mobile manipulation," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023. [Online]. Available: https://arxiv.org/abs/2307.06125

[3] K. Zhou, K. Zheng, C. Pryor, Y. Shen, H. Jin, L. Getoor, and X. E. Wang, "Esc: Exploration with soft commonsense constraints for zero-shot object navigation," in *Proc. International Conference on Machine Learning (ICML)*, 2023. [Online]. Available: https://arxiv.org/abs/2301.13166

[4] M. Ahn, A. Brohan, N. Brown, Y. Chebotar *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," in *Proc. Conference on Robot Learning (CoRL)*, 2022. [Online]. Available: https://arxiv.org/abs/2204.01691

[5] K. Rana, J. Haviland, S. Garg, J. Abou-Chakra, I. Reid, and N. Suenderhauf, "Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning," in *Proc. Conference on Robot Learning (CoRL)*, 2023. [Online]. Available: https://arxiv.org/abs/2307.06135

[6] N. Hughes, Y. Chang, and L. Carlone, "Hydra: A real-time spatial perception system for 3d scene graph construction and optimization," in *Proc. Robotics: Science and Systems (RSS)*, 2022. [Online]. Available: https://arxiv.org/abs/2201.13360

[7] Q. Gu, A. Kuwajima, S. Morin, K. M. Jatavallabhula, B. Sen, A. Agrawal, M. Itkina, A. Stone, K. Chuang, T. Xiao, S. Tulsiani, S. Rusinkiewicz, D. Batra, J. Tenenbaum, A. Torralba, A. Schwing, and K. Fragkiadaki, "Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2024. [Online]. Available: https://arxiv.org/abs/2309.16650

[8] P. Wu, Y. Mu, B. Wu, Y. Hou, J. Ma, S. Zhang, and C. Liu, "Voronav: Voronoi-based zero-shot object navigation with large language model," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2024. [Online]. Available: https://arxiv.org/abs/2401.02695

[9] C. Li, F. Xia, R. Martin-Martin, M. Lingelbach, S. Srivastava, B. Shen, K. Vainio, C. Gokmen, G. Dharan, T. Jain, A. Kullu, S. Ha, R. Vaish, L. Fei-Fei, and S. Savarese, "igibson 2.0: Object-centric simulation for robot

learning of everyday household tasks," *Proc. Conference on Robot Learning (CoRL)*, 2021. [Online]. Available: https://arxiv.org/abs/2108.03272

[10] D. Honerkamp *et al.*, "Moma-llm project page," https://moma-llm.cs.uni-freiburg.de/, 2024.