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

Student Name
Matriculation No: 0000000

Supervisor: Supervisor Name

2026-01-18

Abstract

This report analyzes MoMa-LLM, a novel framework for interactive object search that grounds Large Language Models (LLMs) within dynamically constructed scene graphs [1]. The system addresses the challenge of finding objects that may be hidden inside receptacles or behind closed doors in initially unexplored environments. By extracting structured textual representations from hierarchical scene graphs—incorporating room classifications, object states, and navigation frontiers—MoMa-LLM enables zero-shot, open-vocabulary reasoning for mobile manipulation. The paper introduces AUC-E (Area Under the Efficiency Curve) as a novel evaluation metric. Experiments demonstrate that MoMa-LLM achieves 97.7% success rate, 63.6 SPL, and 87.2 AUC-E in simulation, significantly outperforming HIMOS (93.7% SR, 48.5 SPL) and ESC-Interactive (95.4% SR). Real-world deployment achieves 80% success rate while traveling nearly half the distance of baselines (17.9m vs 33.9m) [1].

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 [3]. 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 [4].
- **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 [5].
- **Other:** Approaches like ConceptGraphs [6] and VoroNav [7] 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 [4].
- **SayCan:** Grounds LLMs in robotic affordances but focuses on fully observed table-top settings [3].

2.3 Object Search

- **ESC:** A baseline for exploration with soft commonsense constraints, scoring frontiers based on object co-occurrences [8].
- **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 with 15 real-world scene scans [9].
- **Robot (Sim):** Fetch mobile manipulator [1].
- **Robot (Real):** Toyota HSR with RGB-D and LiDAR [1].
- **Real Environment:** 4-room apartment (kitchen, living room, hallway, bathroom) with 54 object categories.

3.3.2 Evaluation Metrics

- **Success Rate (SR):** Percentage of episodes where the target is found and `done()` is called.
- **SPL:** Success weighted by Path Length—measures navigation efficiency.
- **AUC-E:** Area Under the Efficiency Curve. *Novel metric* integrating SR across all time budgets. High AUC-E indicates finding objects quickly, not just eventually [1].

3.3.3 Results

Simulation Results

Method	SR (%)	SPL	AUC-E	Avg. Int.
MoMa-LLM	**97.7**	**63.6**	**87.2**	3.9
ESC-Interactive	95.4	62.7	84.5	4.1
HIMOS	93.7	48.5	77.4	4.8
Unstructured LLM	86.3	59.4	77.6	3.6
Random	93.1	50.2	77.0	5.7

Table 1: Simulation results from Table I of the source paper.

Key Insights:

- MoMa-LLM achieves the highest SR (97.7%), SPL (63.6), and AUC-E (87.2) [1].
- Unstructured LLM (raw JSON) achieves only 86.3% SR with 0.41 invalid actions/episode vs 0.19 for MoMa-LLM—structured prompting is essential.
- HIMOS achieves high SR (93.7%) but poor efficiency (SPL 48.5), indicating brute-force exploration.

Real-World Results

- Both MoMa-LLM and ESC achieved 80% SR (8/10 episodes) [1].
- MoMa-LLM traveled 17.9m vs 33.9m for ESC—nearly half the distance.
- MoMa-LLM required 2.2 interactions vs 3.5 for ESC.

4 Discussion

4.1 Strengths

- **Zero-shot Open-Vocabulary Reasoning:** Adapts to novel object/room categories without retraining [1].

- **Structured Grounding:** Scene graphs bridge perception and LLM reasoning, reducing hallucinations (0.19 invalid actions vs 0.41 for unstructured) [1].
- **Real-World Transfer:** 80% success rate with sim-to-real transfer [1].
- **Robustness:** Policy functions even with 27.6% room classification accuracy [1].

4.2 Limitations

- **Perception Dependency:** Requires ground-truth semantic masks, depth, localization, and handle detection [1].
- **Open-Room Layouts:** Door-based separation struggles with open floor plans [1].
- **Computational Latency:** LLM queries dominate runtime; full graph recomputation at each step [1].
- **Sparse Feedback:** Only "success/failure" signals—cannot distinguish gripper slip from locked door [1].

4.3 Future Directions

- **Vision-Language Models:** Replace adjective-based encodings with dense visual representations [1].
- **Noisy Perception:** Construct graphs from raw sensor data without ground-truth assumptions.
- **Holistic Room Clustering:** Incorporate spatial and semantic details beyond door detection.

4.4 Reproducibility

- **Project Page:** https://moma-llm.cs.uni-freiburg.de/ [10].
- **Code:** Referenced but full pipeline details limited.

5 Conclusion

MoMa-LLM establishes a new paradigm for semantic grounding in mobile manipulation by treating the scene graph as a dynamic, structured prompt [1]. The framework demonstrates three key findings:

1. **Structured representations outperform raw data** for LLM-based planning (97.7% vs 86.3% SR).

2. **Interactive search** requires joint reasoning about navigation, exploration, and manipulation—not pairwise scoring.

3. **Dynamic scene graphs** with history realignment enable coherent long-horizon planning.

The empirical validation—particularly the 87.2 AUC-E in simulation and efficient real-world transfer (17.9m vs 33.9m distance)—marks a significant step toward hybrid neuro-symbolic architectures for embodied AI.

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] 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

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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

[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.