Language-Grounded Scene Understanding for Mobile Manipulation

Pratik Adhikari

2026-01-13

Contents

Figures

Tables

Abstract

Large language models (LLMs) have recently been adopted as high-level planners for robotic agents. However, most existing work focuses either on navigation or manipulation in isolation and assumes a fully explored environment. The paper *Language-Grounded Dynamic Scene Graphs for Interactive Object Search with Mobile Manipulation* by Honerkamp *et al.* introduces MoMa-LLM, an approach that uses a dynamically updated scene graph to ground an LLM in partially observable household environments. The proposed system formulates interactive object search as a Partially Observable Markov Decision Process (POMDP) and constructs a two-level scene graph consisting of rooms and objects. It computes a navigational graph from a Bird's-Eye View (BEV) occupancy map using a generalized Voronoi diagram, segments it into rooms via a door-based kernel density estimate and assigns objects to rooms through a distance-weighted path formulation. High-level actions such as *navigate*, *open* and *explore* are mapped to low-level policies and embedded in a structured language prompt that guides the LLM. To evaluate performance, the authors introduce a search efficiency curve and its area under the curve (AUC-E) to capture success as a function of interaction cost. This seminar report critically analyses the problem formulation, algorithmic design and experimental results of MoMa-LLM and discusses its contributions and limitations in the context of mobile manipulation.

1Introduction

Autonomous mobile manipulation robots promise to assist humans in daily life by performing long-horizon tasks such as fetching objects, organising rooms or setting tables. Realising this vision requires agents that can interpret high-level instructions, explore unknown indoor environments, build semantic representations and manipulate articulated objects. Most existing robotics pipelines treat navigation and manipulation separately and rely on a priori maps; moreover, they employ hand-crafted task planners that operate on abstract symbolic representations. Such assumptions break down in the dynamic and cluttered households for which mobile manipulators are intended. Large language models (LLMs) have recently been adopted as high-level planners due to their ability to reason over free-form instructions, but naively integrating them into robotics has proven challenging: LLMs may hallucinate actions that violate physical constraints and often struggle to handle partial observability.

The paper *Language-Grounded Dynamic Scene Graphs for Interactive Object Search with Mobile Manipulation* by Honerkamp *et al.* proposes MoMa-LLM, a system that addresses these issues by grounding an LLM in a structured, dynamically updated scene graph. The problem considered is *semantic interactive object search*: given a language goal such as "bring me the milk from the fridge", the robot must search for the object in an unfamiliar apartment, open doors and drawers, and deliver the item. This task is formulated as a Partially Observable Markov Decision Process (POMDP) with a continuous state space and requires balancing exploration with interaction. MoMa-LLM constructs a two-level scene graph from raw sensor data in real time, combines this representation with a high-level object-centric action space and embeds both into language prompts for the LLM. The system thereby tightly couples reasoning, perception and control.

This report provides a critical analysis of MoMa-LLM. We first survey related work on interactive search, scene graph representations and language-driven robotics. We then summarise the theoretical contributions of the paper, including the mathematical formulation of the dynamic scene graph, the object-to-room assignment and the high-level planning algorithm. Finally, we discuss experimental results and evaluate the strengths and weaknesses of the approach before outlining future research directions.

For completeness we refer the reader to the original publication describing MoMa-LLM[Hon+24] for further details.

2Related Work

Interactive object search sits at the intersection of exploration, semantic mapping, manipulation and language understanding. Classical work on object search focused on enabling robots to navigate through known environments and identify objects using computer vision; these methods often assumed that target objects were visible from afar and ignored the need for manipulation. Recent methods such as Semi-Autonomous Next-Best-Object Selection and RL-based object search introduce learned policies for deciding where to look next, but they still operate on fixed maps and predefined action spaces. In contrast, *interactive* search acknowledges that many everyday objects are hidden behind cabinet doors or inside drawers and therefore requires the robot to interact with the environment.

The use of large language models as planners in robotics has been explored in frameworks such as Reason + ACT (LLM Reasoning Framework) (REACT), which alternates between reasoning and acting via prompts, and Chain-of-Thought planning. Other works embed state information into free-text prompts that instruct the LLM to select discrete actions. These approaches show promise for zero-shot task execution but often neglect the physical constraints of the robot and lack an explicit representation of the environment. MoMa-LLM addresses this by grounding the LLM in a structured scene graph and restricting its action space to high-level primitives, thereby reducing hallucinations.

Semantic scene graphs have emerged as powerful data structures for representing 3D environments. Hydra [HCC22] constructs a multi-layered dynamic scene graph that fuses odometry, mapping and semantic segmentation to enable long-term spatial reasoning. Similarly, semantic mapping systems such as Open Vocabulary Mapping (OVM) and Semantic Graph Memory (SGM) maintain object-level information to support downstream tasks. MoMa-LLM extends these ideas by incorporating a probabilistic door model and an object-to-room assignment that accounts for 3D distances, resulting in a two-level graph tailored to mobile manipulation.

Another relevant line of work focuses on open-vocabulary perception using models such as CLIP, SAM and open-vocabulary object detectors. These models recognise arbitrary classes by leveraging language embeddings and enable the detection of previously unseen objects. MoMa-LLM leverages such detectors to populate its scene graph with object categories and to generate language prompts for the LLM. Finally, prior approaches to language-guided mobile manipulation (e.g., Code as Policies and VAT-LM) have demonstrated the ability of LLMs to generate robot code or symbolic plans. MoMa-LLM differs by combining a learned high-level planner with a rule-based scene representation and explicit action primitives, facilitating interactive search in unknown environments.

3Paper Summary

In this chapter we provide a detailed summary of Honerkamp *et al.*'s MoMa-LLM approach. We begin by describing the problem formulation and then outline how the authors construct a dynamic scene representation. We present the mathematical details of the room segmentation and object assignment procedures, define the high-level action space and discuss how the scene graph is encoded into natural language prompts. Finally, we summarise the evaluation protocol and key results.

1 Problem Formulation

The authors formulate *semantic interactive object search* as a Partially Observable Markov Decision Process (POMDP) $M = (\mathcal{S}, \mathcal{A}, \mathcal{O}, T, P, r)$, where $\mathcal{S}$ is the (continuous) state space describing the robot pose, explored map and locations of objects and doors, $\mathcal{A}$ is a discrete set of high-level actions, $\mathcal{O}$ denotes observations consisting of aligned RGB-D images, $T(s'|s, a)$ and $P(o|s)$ are the transition and observation models, and $r(s, a)$ is the reward function. The agent receives a language goal g such as "find the milk in the fridge" and must decide actions that maximise expected reward while contending with partial observability[33878035504915†L190-L200]. Unlike previous work, the agent must physically open doors and drawers to reveal objects, making the environment non-stationary.

2 Scene Representation

MoMa-LLM builds a hierarchical representation consisting of a Bird's-Eye View (BEV) occupancy map for navigation and a semantic graph describing rooms and objects. At each time step the robot fuses LiDAR and RGB-D observations into a voxel map B_t encoding obstacles and a free-space map F_t. From B_t it constructs an Euclidean signed distance field (ESDF) that assigns to each grid cell the distance to the nearest obstacle. The gradient of the ESDF is used to compute a *generalised Voronoi diagram* (GVD), defined as the set of points with equal clearance to multiple obstacles. The GVD yields a graph $G_V = (V, E)$ whose vertices correspond to Voronoi nodes and whose edges represent traversable corridors[33878035504915†L320-L337]. To ensure connectivity, the authors extract the largest connected component of G_V and sparsify it for efficiency.

2.1 Room Segmentation via Doors

Rather than segmenting rooms at narrow geometric constrictions, MoMa-LLM segments the Voronoi graph using a probabilistic model of door positions. Let

$\{x_i\}_{i=1}^{N_D}$ denote the 2D coordinates of detected door centres. The density of doors is modelled by a kernel density estimate (KDE)

$$\rho_N(x, H) = \frac{1}{N_D} \sum_{i=1}^{N_D} K_H(x - x_i), \tag{3.1}$$

where K_H is a scaled Gaussian kernel and H is the bandwidth matrix[33878035504915†L358-L375]. Edges of G_V that lie within high-probability regions of ρ_N above a threshold are removed, thereby separating G_V into disjoint components corresponding to rooms. The authors choose a bandwidth of 2.0 based on manual tuning[33878035504915†L358-L376]. High-level connectivity between rooms is computed by finding shortest paths in the original graph G_V that traverse exactly two room components.

2.2 Object–Room Assignment

Once the rooms are established, objects detected by the perception module are assigned to rooms by minimising a distance-weighted path. Suppose an object o was observed from viewpoint v_p. Let n_o and n_{vp} be candidate nodes of the room graph corresponding to the object and viewpoint, respectively. The cost of assigning o to n_o is defined as

$$d_w = \min_{n_o, n_{vp} \in G_V^R} \big(\text{path}(n_o, n_{vp}) + d(o, n_o)^\lambda + d(v_p, n_{vp})\big), \tag{3.2}$$

where $\text{path}(n_o, n_{vp})$ is the shortest path length between nodes on the Voronoi graph, $d(\cdot, \cdot)$ is the Euclidean distance, and $\lambda = 1.3$ biases assignments toward nodes close to the object[33878035504915†L384-L402]. Each object is assigned to the room that minimises d_w, preventing cross-wall assignments. Doors may belong to multiple adjacent rooms.

3 High-Level Action Space

MoMa-LLM defines an object-centric high-level action space $\mathcal{A}$ comprising five primitives[33878035504915†L410-L431]:

- **navigate(*room*,*object*)**: navigate to the Voronoi node associated with an object in a specific room using an A-Star Search Algorithm (A*) planner on the BEV map; success is defined as reaching within 1.5 m of the object.
- **go_to_and_open(*room*,*object*)**: navigate to an object and perform an open operation (for doors, the robot moves through the doorway).
- **close(*room*,*object*)**: analogous to the open action but closing.
- **explore(*room*)**: navigate to an unexplored frontier within the room; success is defined as reaching within 0.5 m of the frontier.

Algorithm 1 MoMa-LLM High-Level Interactive Object Search

language goal g, initial scene graph G_S, perception and control modules goal not achieved Acquire depth and RGB observations and update BEV map B_t and free-space map F_t Compute ESDF from B_t and derive Voronoi graph G_V; extract the largest component and sparsify Segment G_V into rooms G_V^R using door KDE Eq. Eq. (3.1) Assign objects to rooms via distance metric Eq. Eq. (3.2) Encode the scene graph and goal g into a structured language prompt Query the LLM with the prompt to obtain a high-level action $a \in \mathcal{A}$ Execute the low-level policy associated with a and update G_S

- **done()**: terminate the episode and evaluate whether the target object has been found.

These primitives restrict the LLM's outputs to feasible behaviours and decouple high-level reasoning from low-level control. Ambiguities arising from multiple instances of the same object class are resolved by selecting the closest instance.

4 Grounded Language Prompting

At each time step the current scene graph and goal are encoded into a structured text prompt for the LLM. The prompt lists each room, its objects and neighbouring rooms, as well as the robot's current room and explored frontiers. Distances and adjacency relations are discretised into qualitative descriptors via thresholding; for instance, the distance between the robot and an object is binned and mapped to adjectives such as "near", "far" or "very far"[33878035504915†L634-L646]. The prompt also provides an action description template and instructs the LLM to output exactly one high-level action. By grounding the language model in an up-to-date graph and restricting its responses, the authors reduce hallucinated actions and improve safety. When the LLM selects an action, a corresponding low-level policy (navigation or manipulation controller) executes it while continuously updating the scene graph.

5 Algorithmic Overview

The overall MoMa-LLM procedure is summarised in Algorithm 1. The agent repeatedly perceives the environment, updates its scene graph, queries the LLM for an action and executes the corresponding low-level policy until the goal is achieved.

6 Evaluation and Results

The authors evaluate MoMa-LLM in simulation using the iGibson2 environment and in a real-world apartment. They compare against a reinforcement-learning

(RL) baseline trained to follow language instructions, a hierarchical RL baseline and two ablations: an unstructured LLM that receives unorganised scene information and a variant without distance encoding. Success rates, path length and the number of infeasible actions are reported. Traditional metrics such as Success weighted by Path Length (SPL) summarise performance at a fixed budget but ignore the costs of interactions; therefore the authors propose an *efficiency curve* that plots the fraction of episodes solved as a function of the number of low-level steps and define its area under the curve (AUC-E) as a summary[33878035504915†L626-L647]. A perfect policy that finds all objects in a single step achieves AUC-E = 1, while a policy that never finds an object yields zero. When calculating the efficiency curve, the authors weigh each open or close interaction as costing 30 time steps to reflect a real-time duration of roughly 30 s[33878035504915†L1178-L1184]. Experiments demonstrate that MoMa-LLM significantly outperforms baselines in both success rate and AUC-E; the ablation without distances suffers more room assignment errors and requires longer paths, highlighting the importance of the distance-weighted assignment.

4Discussion

The MoMa-LLM framework represents a significant step toward language-guided mobile manipulation in unknown environments. By combining a dynamic scene graph with a structured action space and a grounded large language model, the authors demonstrate successful interactive object search and manipulation. In this discussion we evaluate the contributions of the paper, highlight its strengths and point out several limitations and avenues for future research.

1 Strengths and Contributions

Structured grounding reduces hallucinations. A key insight of MoMa-LLM is that LLMs benefit from explicit grounding in a structured representation. By constraining the action space to a small set of high-level primitives and encoding only relevant rooms, objects and distances, the LLM produces fewer infeasible commands compared to unstructured prompts. The distance-weighted room assignment further ensures that the robot navigates through the correct door rather than driving into walls. These design choices translate into higher success rates and efficiency as evidenced by the reported AUC-E scores[33878035504915†L626-L647].

Integration of navigation and manipulation. Unlike prior work that treats exploration and interaction separately, MoMa-LLM operates within a Partially Observable Markov Decision Process (POMDP) framework that couples mapping, room segmentation and object manipulation. The hierarchical scene graph allows the planner to reason at multiple granularities (rooms and objects), and the high-level actions interleave navigation, opening/closing and exploration. This integrated perspective enables the robot to find hidden objects that require opening drawers or cabinets, a capability missing in many language-guided navigation methods.

Clear evaluation and new metric. The authors recognise that existing metrics such as SPL depend on arbitrary budgets and neglect interaction costs. Their proposed efficiency curve and AUC-E capture performance across a range of budgets and emphasise the time cost of opening or closing objects[33878035504915†L626-L647]. Reporting both simulation and real-world experiments further strengthens the validity of the results.

2 Limitations and Future Work

Reliance on accurate perception. The system assumes access to reliable open-vocabulary object detection, semantic segmentation of rooms and precise pose estimation. In

practice, perceptual errors and occlusions can degrade the quality of the scene graph. Future work could incorporate probabilistic representations and uncertainty propagation to make the planner robust to perception noise. Additionally, the door KDE requires manually chosen bandwidth and threshold parameters, which may not generalise to new environments.

LLM reasoning remains opaque. While grounding the language model reduces hallucinations, the LLM is still treated as a black-box planner. It is difficult to guarantee that the LLM will always select safe and efficient actions, especially when presented with unseen scenarios. Recent work on verifiable planning and neuro-symbolic reasoning could be integrated to provide formal guarantees. Alternatively, the LLM could be trained or fine-tuned with demonstrations to bias it toward desirable behaviours.

Scalability and computational cost. Constructing ESDFs and Voronoi diagrams at every time step can be computationally expensive, particularly in large environments. Sparsification mitigates some of this cost, but maintaining a high-fidelity map and scene graph may not scale to multi-floor buildings. Moreover, each LLM query incurs latency and requires internet connectivity (for hosted models). Exploring lightweight language models or localised planning heuristics might alleviate these issues.

Limited generalisation. The evaluation focuses on indoor household environments with a relatively small set of rooms and objects. Although the system demonstrates impressive performance in this domain, its applicability to outdoor environments, cluttered warehouses or industrial settings remains unclear. Extending the scene representation to incorporate 3D geometry, dynamics and affordances could broaden its applicability.

In summary, MoMa-LLM effectively integrates scene graph reasoning with language-driven planning to achieve interactive object search. Addressing the limitations identified above will be crucial for deploying such systems in real-world applications at scale.

5Conclusion

This seminar report analysed the MoMa-LLM system for language-grounded interactive object search with mobile manipulation. We described how the authors formulate the task as a Partially Observable Markov Decision Process (POMDP), construct a dynamic two-level scene graph using Voronoi diagrams and a door-based kernel density estimator, assign objects to rooms via a distance-weighted cost and define a concise high-level action space. We examined the structured language prompts that ground a large language model in this scene representation and summarised the evaluation using a new efficiency curve metric.

The key contribution of MoMa-LLM lies in its tight coupling of perception, mapping, language understanding and control. By embedding an LLM within a principled representation of the world, the system achieves higher success rates and search efficiency than reinforcement learning baselines and unstructured prompting. At the same time, the method relies on accurate perception, manual parameter choices and black-box reasoning. Addressing these limitations—through robust perceptual models, verifiable planning and more scalable graph abstractions—presents an exciting avenue for future research.

Overall, MoMa-LLM demonstrates the potential of combining dynamic scene graphs with large language models to enable robots to perform complex tasks in unexplored environments. The insights gained from this work will be valuable for advancing mobile manipulation toward real-world deployment.

AAppendix

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Bibliography

[HCC22] N. Hughes, Y. Chang, and L. Carlone. "Hydra: A Real-time Spatial Perception System for Visual-Inertial Agents". In: *Robotics: Science and Systems (RSS)*. 2022.

[Hon+24] D. Honerkamp, M. Büchner, F. Despinoy, T. Welschehold, and A. Valada. "Language-Grounded Dynamic Scene Graphs for Interactive Object Search With Mobile Manipulation". In: *IEEE Robotics and Automation Letters* 9.10 (Oct. 2024), pp. 8298–8305. ISSN: 2377-3766. DOI: 10.1109/LRA.2024.3441495.