ChatGPT

# Evaluating MoMa-LLM's Contributions in Robotics

MoMa-LLM (Language-Grounded Dynamic Scene Graphs for Interactive Object Search with Mobile Manipulation) introduces several claimed contributions. Below, we break down each contribution and **critically assess what is truly novel** about it in the context of prior robotics research, rather than taking the authors' claims at face value.

## Dynamic 3D Scene Graph as a World Model (Hybrid Representation)

**Claimed Contribution:** *"A scalable scene representation centered around a dynamic scene graph with open-vocabulary room clustering and classification."* [1] In MoMa-LLM, the robot builds a **3D dynamic scene graph** of its environment as it explores, augmented with an open-vocabulary semantic understanding of rooms (clustering spaces into 'kitchen', 'bedroom', etc. on the fly). This scene graph is coupled with a navigational **Voronoi graph** for path planning, creating a hybrid world model for long-horizon tasks [2].

**Assessment:** *Is a dynamic 3D scene graph itself a new contribution?* 3D scene graphs have been studied before in robotics and vision. Prior work established scene graphs as a hierarchical representation of environments (e.g. nodes for objects/rooms and edges for relations) for high-level reasoning [3]. Dynamic or **incrementally-built** scene graphs also existed – for example, Hydra (RSS 2022) introduced a real-time spatial perception system for dynamic 3D scene graph construction [4] [5]. **However, MoMa-LLM's novelty lies in how it uses this representation in combination with other elements, not in the mere existence of a scene graph.** Specifically, MoMa-LLM is the first to leverage a *dynamic* scene graph as the central world model for an **interactive mobile manipulation task** (object search in an unknown environment) while using *open-vocabulary* semantics. Earlier robotics works either used static, fully-known scene graphs or did not incorporate real-time semantic expansion. For instance, ConceptGraphs (ICRA 2024) and VoroNav (2024) explored open-vocabulary scene graphs and zero-shot perception for planning, but they did not address **interactive object search in a changing environment** [6]. In fact, none of the prior works combined a **continuously-updated** scene graph with an **interactive** task (where the robot opens doors, drawers, etc.) – *"realizing object navigation using both dynamic and interactive scene graphs has not been tackled thus far"* in those works [7].

Moreover, MoMa-LLM's scene graph includes on-the-fly **room clustering and labeling** (the robot infers room regions and labels them like "kitchen" based on objects inside). This open-vocabulary room classification is uncommon; previous systems typically assumed a fixed set of room types or required a prior map. By integrating mapping, object detection, semantic labeling, and graph updates during exploration, MoMa-LLM **brings together modules that were previously separate**. This integration – a unified **spatial-semantic world model** for an unexplored home – is a key part of their contribution. While 3D scene graphs themselves aren't new, MoMa-LLM is unique in *how* it employs them: as a live, structured memory that a language model can query for planning in a **long-horizon, multi-room task** [6]. No prior single system had achieved this specific synthesis (open-vocab semantics + dynamic updates + interactive use) in robotics.

## Grounding LLM Planning in Structured Scene Knowledge

**Claimed Contribution:** *"Structured compact knowledge extraction to ground LLMs in scene graphs for large unexplored environments."* [8] In simpler terms, MoMa-LLM uses the scene graph to **generate a concise textual description of the world state**, which is fed into a Large Language Model. The LLM then acts as a high-level decision-maker (policy), choosing the next action for the robot from a predefined set, based on this structured state [9] [10]. This grounds the LLM's abstract reasoning in the robot's current reality.

**Assessment:** *Is using an LLM for high-level planning with scene graph context novel?* The idea of an LLM guiding a robot is emerging, but MoMa-LLM is among the first to do so in *unexplored, multi-room environments*. Previously, LLM-based planners were tested mostly in fully-known or small-scale settings. For example, *SayCan* (Google, 2022) used an LLM to suggest actions, but it only filtered those suggestions with affordance scores and **did not utilize any spatial-semantic map** of the environment [11]. Other works like LLM-Planner (ICCV 2023) and SayPlan (CoRL 2023) incorporated scene information, but under much simpler conditions: either by retrieving similar known scenarios or by assuming a **complete scene graph is given in advance** [12]. In contrast, MoMa-LLM operates **zero-shot** in a **partially observable, growing environment**, which is far more challenging.

Before MoMa-LLM, if a robot was to search for an object, methods either relied on learned policies (e.g. reinforcement learning over maps) or simple heuristics using semantic cues. Those that did use language models provided the LLM with relatively unstructured or partial information – e.g. a *list of observed objects* so far, or a raw JSON of the scene graph [13]. The MoMa-LLM authors argue that such *"simple prompting strategies, such as lists of observed objects or raw JSON input of a full scene graph, become insufficient"* for large, complex scenes [13]. Instead, MoMa-LLM **extracts only the relevant knowledge** from the scene graph (for instance, the rooms discovered and what key objects are in each room, which doors are closed or open, etc.) and feeds that in a structured natural-language format to the LLM [14]. This compact grounding prevents the LLM from hallucinating plans that defy the current state of the world, by *"guiding the LLM to adhere to the physical realities of the scene"* [14].

Critically, **no prior work had tied an LLM's decision-making this tightly to a dynamically built scene representation in an unknown environment**. Some researchers fine-tuned smaller language models on scene graphs (e.g. Chalvatzaki et al. 2023 finetuned GPT-2 on structured scene data [12]), but that was for *fully known* scenes rather than live exploration. Another recent system, *SayNav (ICAPS 2024)*, is conceptually closest to MoMa-LLM: it also uses an LLM with a scene graph for navigation. However, SayNav was limited to non-interactive tasks in single rooms and assumed certain knowledge (it restricted the LLM to a single room's subgraph and even hard-coded when to move to the next room through an open door) [15]. MoMa-LLM removes these restrictions – the LLM sees the whole discovered map, can decide to search another room or open a closed door autonomously, and it's all **open-vocabulary** (no fixed list of object or room types) [15]. In summary, MoMa-LLM's second contribution is **showing that an LLM can serve as a high-level planner grounded in a robot's internal scene graph**, which *had not been demonstrated before in prior robotics research* in this form. The approach is novel in that it bridges powerful pre-trained knowledge (the LLM's "common sense" about which objects are likely in a kitchen, etc.) with the robot's live mapped knowledge, yielding plans that are both informed by human-world knowledge and **anchored to the robot's current environment state** [16] [17].

*Why hadn't others done this?* Large Language Models controlling robots is a very recent trend (post-2022). Early attempts struggled with grounding – LLMs would suggest actions that weren't possible in the actual environment. MoMa-LLM is among the first to solve this by giving the LLM a structured description of what **has been seen and what remains unknown**. Previous robotics systems

either lacked such a rich world model or didn't integrate it with an LLM. Thus, MoMa-LLM's contribution here is **in the integration**: it combines mapping, perception, and language reasoning in a way that wasn't explored before, enabling zero-shot high-level planning in complex, novel environments [18] [19].

## Semantic Interactive Object Search Task

**Claimed Contribution:** *"Semantic interactive search task for large scenes with numerous objects and receptacles."* [20]. MoMa-LLM introduces a new evaluation scenario: the robot must find a specific target object in a realistic indoor environment, which may require **opening doors, cabinets, or drawers** and searching multiple rooms. Crucially, the environment preserves real-world **semantic object distributions** – objects are placed where you'd expect in a home (e.g. a toothbrush in a bathroom, plates in a kitchen cabinet), rather than random locations.

**Assessment:** *Is this task truly new in robotics?* It builds on prior object search tasks, but extends them. **Interactive object search** (where a robot can physically open or move obstacles to find things) was first formalized by Schmalstieg et al. (RA-L 2023) [21]. In that earlier task, the robot had to search for objects hidden behind closed doors or inside containers, but importantly, the object placements were **random** and the test environments had a limited number of object types and receptacles [22]. MoMa-LLM's task variant adds the *semantic* element: rather than random placements, the target object is one of the many objects naturally present in the environment and is likely to be found in a semantically appropriate location. As the authors state, *"we introduce a semantic single-object search variation of [the interactive search] task, which uses all objects in the scene and keeps the semantic co-occurrences in the scene intact."* [23]. In other words, **the entire apartment is populated with objects in a realistic way**, and any of those could be the search target – this makes it a larger-scale and more life-like challenge than earlier benchmarks.

No previous benchmark in robotics has combined **large-scale indoor exploration** with **semantic reasoning and physical interaction** to this degree. Before, you either had: (1) **ObjectGoal navigation** tasks (find an object category in a house) which involve semantic reasoning but usually assume doors are open or no articulated obstacles, *or* (2) **Interactive search** tasks which involved opening doors/drawers but on a smaller scale or without leveraging semantic knowledge (since the target was random, the robot couldn't use prior knowledge of likely locations). MoMa-LLM's task merges these: the robot must use common-sense knowledge (e.g. knowing mugs are probably in the kitchen) *and* perform multi-step interactions to actually look in the right places (open the kitchen cabinets). This is indeed a **new combination** of requirements. The authors also emphasize that unlike *non-interactive* semantic navigation methods which might predict a likely room or direction for the target [24], their scenario demands **reasoning over multiple steps** – the agent might have to navigate to the correct room, then interact (open a door, then a cabinet) to retrieve an object [25]. This kind of *long-horizon, semantic and physical* search problem had not been concretely benchmarked before.

So, while the building blocks (semantic object search, interactive exploration) have been studied separately, **MoMa-LLM's contribution is defining a unified task that is more complex and realistic** than each alone. It challenges a robot to leverage human-like understanding of environments in service of a physical goal. This task serves as a platform to evaluate MoMa-LLM and compare it to prior methods on a more **robust measure of real-world performance** – because a method that simply wanders randomly or relies on brute-force would fare poorly when there are many rooms and most are irrelevant to the query. By introducing this benchmark, the authors fill a gap: as they note, previous works either dealt with interactive search *without* semantic context or semantic search *without* interactive obstacles [23] [26]. MoMa-LLM required and thereby *pioneered* a task that needs both.

*(As an aside, commercial robotics is still far from this level of autonomous object search. Current household robot products do not yet combine semantic understanding and interactive searching in general deployment – those are mostly research prototypes. For instance, demos like Google's Palm-SayCan or Microsoft's ChatGPT-controlled robot show robots following language instructions, but their environments are often preset or single-room and they don't build a rich scene graph. MoMa-LLM's task is a step toward what a home-assistant robot might need to do: "Find my keys" even if they are in another room and inside a drawer, which no off-the-shelf system can reliably handle today.)*

## New Evaluation Paradigm: Efficiency Curves (AUC-E)

**Claimed Contribution:** Along with the task, the paper proposes a *"novel evaluation paradigm for object search tasks through full efficiency curves, instead of a single time budget"* [20]. They introduce **Efficiency Curves** that plot the success rate of the robot as a function of time (or steps) taken, and define **AUC-E (Area Under the Curve of Efficiency)** as a single metric summarizing how quickly and reliably the agent finds the object [27].

**Assessment:** *Why is a new metric needed, and is it new?* In prior object-search evaluations, it was common to use a binary success rate given a fixed budget (e.g. did the robot find the object within 1000 steps or 5 minutes). The cutoff was arbitrary – a method might barely fail at 1000 steps, counting as failure, while another finds it at 999 steps and counts as success. That can skew comparisons. MoMa-LLM's evaluation instead looks at **the whole curve**: for any time threshold, what fraction of trials succeeded by then. This gives a more nuanced picture of performance over time. By integrating that curve, AUC-E captures both *whether* and *how fast* the agent succeeds in one number [28] [29]. This approach is analogous to evaluating algorithms on multiple budgets to avoid biasing to one cutoff.

This **is a novel contribution in the robotics object-search domain**. To our knowledge, earlier works did not use the area-under-curve of success rates as a metric for object search; they typically reported success at one or a few fixed time limits [27]. The MoMa-LLM authors argue this gives a fairer comparison, removing dependency on an "arbitrary time budget" [27]. It's a logical improvement – similar in spirit to how detection algorithms report precision-recall curves rather than a single threshold of IoU. By introducing AUC-E, they encourage future researchers to consider the *efficiency* of search strategies, not just eventual success. In summary, while defining a new metric is a more minor contribution than the core algorithm, it is still significant because it **changes how results are measured** for this class of problems. It makes the evaluation of long-horizon tasks more **comprehensive** (considering the entire timeline of the attempt) and thus is a welcome contribution for benchmarking.

(We should note that the task+metric go hand-in-hand: the complexity of the semantic search task means that efficiency truly matters, so evaluating by a curve is very appropriate. In a trivial task, success is usually achieved quickly or not at all, but in MoMa-LLM's scenario, one method might find the object in 2 minutes vs another in 5 minutes, which the AUC-E captures as a meaningful difference.)

## Conclusion: What is Unique about MoMa-LLM?

MoMa-LLM brings together ideas from mapping, semantic scene understanding, and language-model planning in a way that **no prior robotics system had done**. The key novelty is not any single ingredient like "scene graphs" or "LLMs" in isolation – those existed – but the *integration* of these into a coherent system for a challenging real-world task. In particular, MoMa-LLM is the first to **ground a large language model in a continually-updated 3D scene graph of an unexplored environment** to drive an interactive object search. This required: a) a hybrid world representation (metric + semantic) that

scales to large homes, b) translating that world state into structured language for the LLM, and c) leveraging the LLM's general knowledge to guide exploration efficiently. **No one had combined dynamic scene graphs with open-vocabulary semantics and LLM-driven planning for long-horizon tasks before** [6] [15]. Additionally, MoMa-LLM introduced a more realistic search benchmark and a better way to evaluate performance (efficiency curves), which together push the field toward more **comprehensive and grounded evaluations** [23] [27].

In summary, while concepts like 3D scene graphs or semantic search were known, MoMa-LLM's **specific contributions** are unique in robotics: it **integrates multiple modules** (mapping, vision, semantics, and language reasoning) into one system and demonstrates clear benefits over prior methods in a complex task [30] [15]. By doing so, it has opened up a new avenue for robots that can reason like **LLMs** yet act grounded in a real physical world model – something that had not been shown before MoMa-LLM. The paper's contributions, especially the grounding of LLM planning in an interactive scene graph, are therefore *substantive novelties* in the domain, not just reusing old ideas. Each contribution addresses a gap left by previous approaches (lack of long-term memory, lack of semantic reasoning in interactive tasks, or ad-hoc evaluation), making MoMa-LLM a noteworthy step forward in mobile manipulation research.

**Sources:**

- Honerkamp et al., *"Language-Grounded Dynamic Scene Graphs for Interactive Object Search with Mobile Manipulation (MoMa-LLM),"* IEEE RA-L 2024 – Introduction and Contributions [16] [6] [31] [23], and comparative discussions in Related Work [30] [15]. These detail the novelty of dynamic & interactive scene graphs, LLM grounding, the semantic search task, and the new AUC-E evaluation.
- Schmalstieg et al., *"Learning Hierarchical Interactive Multi-Object Search for Mobile Manipulation,"* RA-L 2023 – introduced the baseline interactive search task (with random placements) that MoMa-LLM builds upon [22]. This highlights how MoMa-LLM's task differs by incorporating semantic realism.
- Related prior works like SayCan (Google, 2022) and SayNav (2024) as cited in Honerkamp *et al.* – used LLMs or scene graphs in isolation, but had limitations (single-room, no dynamic updates, etc.) [32] [15], underscoring MoMa-LLM's integrated approach as a first in the field.

---

[1] [3] [4] [5] [6] [7] [8] [12] [13] [14] [15] [16] [17] [18] [19] [20] [21] [22] [23] [24] [25] [26] [27] [28] [29] [30] [31] [32] Honerkamp2024 - Language Grounded Dynamic Scene Graphs for Interactive Object Search with Mobile Manipulation.pdf
file://file_0000000041c071f4b7c082f4d595d8fd

[2] [9] [10] [11] main.pdf
file://file_0000000099ac71f4b5e2d2684bb35a20