LOOKPLANGRAPH: EMBODIED INSTRUCTION FOLLOWING METHOD WITH VLM GRAPH AUGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, approaches using Large Language Models (LLM) as planners for robotic tasks have become widespread. In such systems, the LLM must be grounded in the environment in which the robot is operating in order to successfully complete tasks. One way to do this is to use a scene graph that contains all the information necessary to complete the task, including the presence and location of objects. In this paper, we propose an approach that works with a scene graph containing only immobile static objects, and augments the scene graph with the necessary movable objects during instruction following using a visual language model and an image from the agent's camera. We conduct thorough experiments on the SayPlan Office, BEHAVIOR-1K, and VirtualHome RobotHow datasets, and demonstrate that the proposed approach effectively handles the task, bypassing approaches that use pre-created scene graphs.

1 INTRODUCTION

The pursuit of autonomous agents that can comprehend and execute complex human instructions in dynamic environments is a fundamental objective in robotics. Recent strides in Large Language Models (LLMs) have shown significant potential in reasoning and planning for a variety of tasks articulated in natural language (Huang et al., 2022; Ahn et al., 2022; Singh et al., 2023). For robots to effectively carry out these tasks, it is crucial that LLMs are grounded in the physical environments where the robots operate. One effective strategy for achieving this grounding is through the use of scene graphs (Gu et al., 2023), which offer structured representations of environments by detailing objects and their interrelationships.

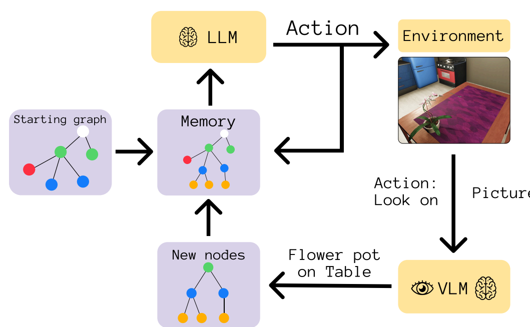

Figure 1: **LookPlanGraph** enhances an agent's ability to operate in dynamic environments by integrating real-time updates from the environment into its graph representation.

Traditionally, the processes of constructing a scene graph and executing tasks using it have been treated separately. SayPlan (Rana et al., 2023) leverage static scene graph representations to generate viable task plans for embodied agents. However, this reliance on static graphs presupposes unchanging environments, a condition rarely met in real-world scenarios where objects frequently change locations or states. Consequently, when the environment undergoes changes, methods such as SayPlan require the entire scene graph to be reconstructed. This reconstruction involves additional procedures like scene navigation, image capturing, and data analysis, all of which are time-intensive and computationally demanding, thus impeding real-time application.

The assumption of a static scene graph is particularly impractical in dynamic settings for several reasons. Firstly, other agents or unforeseen events may alter the state, location, or relationships of objects within the environment. Secondly, certain objects might be concealed within closed containers like boxes or cabinets. Including these hidden objects in the initial graph would necessitate a thorough examination of all possible storage spaces during graph construction, an approach that is neither efficient nor scalable. Therefore, an agent operating in a dynamic environment must possess the capability to dynamically extend and update its scene graph based on real-time observations made during task execution.

Another significant limitation of existing methods like SayPlan is their dependency on large, computationally intensive models such as GPT-4. While these models are powerful, their substantial resource requirements pose challenges for applications that necessitate localized computation or operate under hardware constraints. To overcome these challenges, we present three key contributions:

1)Development of SayPlan Lite: We present SayPlan Lite, a streamlined version of the original SayPlan method, designed to boost the efficiency of smaller LLMs for local machine use. Its success showcases the potential for broader application in resource-limited contexts, making it a viable tool for building LLM agents tailored to constrained environments.

2)Proposal of LookPlanGraph for Dynamic Environments: We propose LookPlanGraph, a graph-based planning framework for dynamic environments. Unlike static scene graphs, this approach initializes with unmovable assets and dynamically updates with movable objects using a Visual Language Model (VLM) and the agent's egocentric camera. It employs a Memory Graph Mechanism to adapt to environmental changes by focusing on relevant, nearby objects, reducing computational demand. A Graph Augmentation Mechanism further allows real-time exploration and updates, ensuring adaptability to the agent's surroundings.

3)Compilation of a Graph Dataset: We have created a 558-task dataset for graph-based instruction-following, featuring automated validation. Built from SayPlan Office, BEHAVIOR-1K, and VirtualHome RobotHow environments, this dataset offers a robust resource for assessing planning methods across diverse settings.

2 Related works

2.1 Embodied Planning

Robotic task planning generates sequences of actions to achieve goals within an environment. Traditional methods use domain-specific languages, such as PDDL (Fox & Long, 2003) and Temporal Logic (TL) (Doherty & Kvarnstram, 2001), combined with parsing, search methods, and heuristics. These are effective in controlled settings but struggle with scalability and generality in complex environments. Recently, LLMs are being utilized for task planning due to their in-context learning abilities. Huang et al. (2022) used LLMs to translate actions into executable commands specific to environments. LOTA-Bench (Choi et al., 2024) employs LLMs to predict the next action based on sequence probability, while the LLM+P approach (Liu et al., 2023) integrates grounding by creating a PDDL description for classical planners.

Effective grounding needs a reliable environmental representation, achieved by scene graphs, which organize entities and their relationships (Gu et al., 2023; Liu et al., 2021; Devarakonda et al., 2024). Innovative methods like Delta (Liu et al., 2024b) extend the LLM+P approach using scene graphs to describe PDDL domains. KARMA (Wang et al., 2024) introduces a memory-augmented system for LLM-based planning in embodied AI agents, integrating long-term 3D scene graphs and dynamic short-term memory. Dai et al. (2024) integrates LLMs with hierarchical metric-semantic models for task planning over scene graphs. It translates natural language tasks into LTL automata and introduces an optimal hierarchical planning method guided by LLM heuristics. SayPlan (Rana et al., 2023) prompts scene graphs in LLMs to make methods executable. These methods effectively combine scene graphs with LLMs for task planning, particularly with large models. However, smaller models face challenges due to limited capacity and context handling, highlighting the need for optimized solutions for smaller architectures.

2.2 VLM integration

Recent VLM advancements benefit robotics by interpreting visual data directly. Unlike LLM, VLM handle raw pixel data, enhancing robotic perception. ActPlan-1K (Su et al., 2024) introduces a benchmark to evaluate VLM in multi-modal and counterfactual task planning, revealing their limitations in generating effective procedural plans. RT-2 (Brohan et al., 2023) and PaLM-E (Driess et al., 2023) integrate VLM for multimodal inputs in robotics. ViLa (Hu et al., 2023) employs VLM

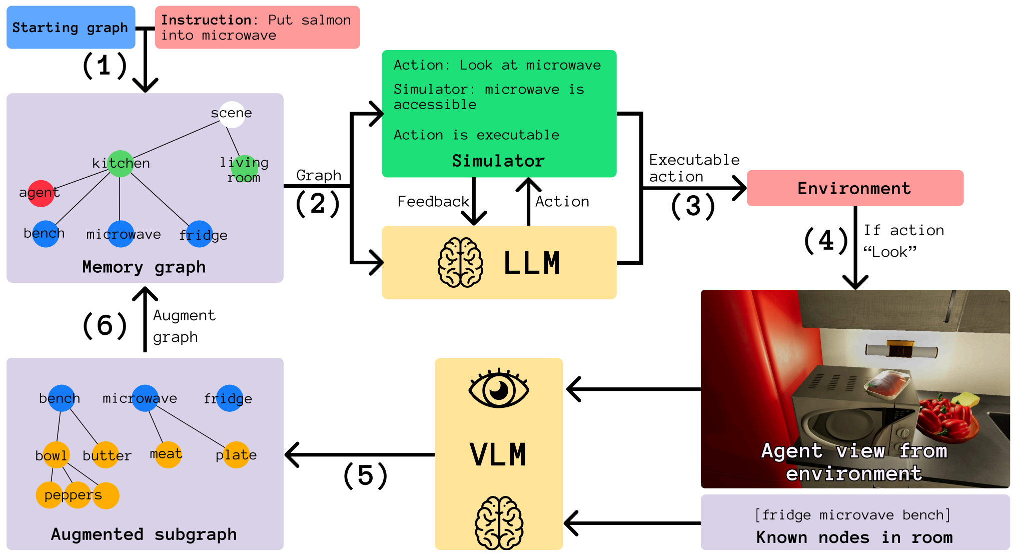

Figure 2: **LookPlanGraph Overview:** The LookPlanGraph starts with an instruction and a static environment graph **(1)**. A memory graph, initially a copy of the starting graph, is processed with the task description by the LLM and is also sent to the Scene Graph Simulator **(2)**. The LLM suggests an action, which the Simulator checks for feasibility. If executable, the action changes the environment, updating the memory graph **(3)**. For actions needing visual feedback (e.g., "look"), the environment sends a camera view to the VLM **(4)**. The VLM processes this image, along with room node data from the memory graph, to produce an augmented subgraph **(5)**, which updates the memory graph **(6)**. This cycle **(2-6)** repeats until the LLM decides the task is complete.

for direct task execution. The integration of VLM with graph structures shows promise due to VLMs ability to build and use graph-based representations. ConceptGraphs (Gu et al., 2023) uses 2D model outputs to create 3D scene graphs and generate plans using LLMs. VeriGraph (Ekpo et al., 2024) emphasizes benefits of combining graphs with VLM, especially in complex tasks. While often in smaller settings, structured representations like graphs can greatly improve planning efficiency in complex scenarios.

3 Problem Formulation

We address the challenge of enabling an autonomous mobile manipulator robot to plan, navigate, and manipulate objects within large-scale household environments using natural language instructions. The robot must reason about dynamic scenes and adapt to environmental changes, such as object locations and states. Our solution necessitates generating executable actions that involve complex navigation and manipulation tasks, effectively accommodating the dynamics of multi-room environments where static representations are inadequate.

Formally, given a 3DSG G and a task instruction T expressed in natural language, our framework, LookGraphPlan, can be conceptualized as a high-level decision-making module denoted by $\pi(\mathbf{a} \mid T, G)$. This decision-making module is capable of generating an action $\mathbf{a}$ that is grounded in the environment where the embodied agent operates. Furthermore, the generated action is designed to concurrently follow the instruction to be carried out.

Our approach focuses on two primary challenges: 1) Develop a framework capable of operating with scene representations that initially include only the static elements of the environment. This framework must enable the agent to iteratively expand and modify the scene graph during task execution, reflecting changes in the environment; 2) Enabling effective planning using smaller LLMs, that can run locally, to reduce reliance on computationally intensive, large-scale models.

4 METHOD

We introduce LookPlanGraph, a scalable method leveraging scene graphs to operate in environments without predefined states or positions for movable objects. Central to this approach is the Scene Memory Graph (SMG), which is continuously maintained and updated to reflect the current state of the environment. This method empowers exploration and interaction within the environment, augmenting the SMG with newly detected objects.

Illustrated in Figure 2, the LookPlanGraph methodology integrates a LLM, a scene graph representation, and VLM to execute tasks. The process begins with an starting graph outlining the rooms and fixed assets within the environment, which is then replicated into the SMG for use throughout the method. As the agent engages with the environment, the SMG is dynamically updated, incorporating modifications made by the agent via the Scene Graph Simulator and new objects identified by the VLM. Algorithm 1 of the method follows a structured cycle (4-15):

Algorithm 1 LookPlanGraph

1: **Given:** LLM planner *LLM*, VLM parser *VLM*, Environment *ENV*, Memory graph M, Graph Simulator Sim
2: **Inputs:** Starting graph *G*, Task *T*
3: $M = G$
4: **while** action $!=$ done **do**
5:     **while** feedback $!=$ None **do**
6:         action $\leftarrow LLM(M, T, feedback)$
7:         feedback $\leftarrow Sim(M, action)$
8:     **end while**
9:     *ENV*(action)
10:     **if** action$='look_on'$ **then**
11:         new nodes $\leftarrow$ *VLM*(*ENV*, M)
12:         M.append(new nodes)
13:     **end if**
14: **end while**

LLM Decision-Making (7): The SMG is encoded into a prompt and provided to the LLM, along with the task instructions. Using this input, the LLM determines the next action for the agent to perform. In our approach, the list of possible actions is limited to manipulation tasks such as *goto, pick up, open, close, put on, put in,* and two scene exploration actions: *look on* and *look inside*.

Simulation and Feedback (5-9): The proposed action is sent to the Scene Graph Simulator, which evaluates its feasibility. If the action is valid, the simulator updates the SMG to reflect the outcome. If the action is invalid, feedback is returned to the LLM for re-planning.

Environment Interaction (10): Once validated by the simulator, the action is executed in the real environment.

Graph Augmentation via VLM (11-14): For exploration actions, such as *look on* or *look inside*, the VLM is invoked. The VLM processes images of the environment and generates nodes for newly identified objects. These objects are then added to the SMG.

4.1 MEMORY GRAPH

The Memory Graph (Figure 3) is a hierarchical, graph-based structure inspired by 3D Scene Graphs (Kim et al., 2019; Kurenkov et al., 2021). It encodes spatial semantics, object relationships, and affordances for efficient robotic planning (Gay et al., 2019; Rosinol et al., 2021). Organized into four layers—Scene, Place, Asset, and Object—it abstracts the environment at different levels. The Scene Layer represents the entire environment, the Place Layer defines areas (e.g., rooms), the Asset Layer includes immovable objects, and the Object Layer contains movable items. An additional agent node tracks the agent's position and interactions.

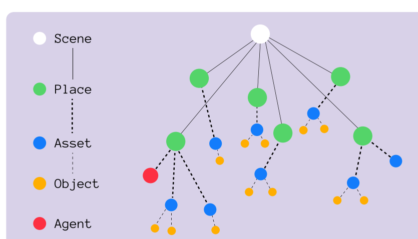

Figure 3: Memory graph structure constructed with four layers and an additional agent node to track the agent's position.

4.2 ACTION GENERATION

Action generation by the LLM involves structured prompting, graph filtering, and a feedback loop with the Scene Simulator. The prompt consists of three parts: a static prompt, dynamic components, and feedback. The static prompt describes actions and states derived from the graph, focusing on essential home environment tasks like pick-and-place, opening/closing, and turning devices on/off.

Additional actions, such as “look on” and “look inside”, enable visual interaction. A detailed prompt structure and the full action list are provided in Appendix A.

To optimize for compact LLM models, the prompt is concise yet informative by including only objects in the agent’s immediate environment (e.g., the same room) and adding previously interacted objects, especially for long-term planning tasks. The Scene Graph Simulator, similar to SayPlan, ensures actions are executable in the real environment. It attempts to execute LLM-generated actions and updates the memory graph and environment state if successful. If an action fails due to constraints, the simulator provides feedback, which is included in the next prompt to improve subsequent actions.

4.3 Graph augmentation

For visual interaction with the environment, the LLM can inspect the top or interior of accessible assets by calling corresponding action and incorporate newly discovered objects into the memory graph. Example of such interaction shown in the right part of Figure 2. This process involves capturing an image from the environment that represents the agent’s field of view. The image is then passed to a Vision-Language Model along with list of assets and objects already present in the memory graph to ensure that only new nodes are added. The VLM is prompted to identify new objects, their states, and their relationships with existing assets. Subsequently, the identified nodes are integrated into the memory graph, making them available for future interactions.

5 Dataset preparation

Graph-based scene representation methods are gaining traction in research, yet there is a noticeable lack of 3DSG datasets tied to specific tasks, which has led researchers to rely on proprietary data. To bridge this gap, we have curated a comprehensive dataset by integrating resources from multiple sources. Specifically, we combined textual instructions and environmental data from BEHAVIOR-1K (Li et al., 2024), VirtualHome RobotHow (Puig et al., 2018), and SayPlan Office (Rana et al., 2023). While BEHAVIOR-1K lacks task descriptions and graph representations, VirtualHome RobotHow offers graph representations but not in the 3DSG format, and SayPlan Office lacks a coded implementation. Our curated dataset addresses these limitations, providing a valuable resource for evaluating and advancing graph-based methods in robotics research.

Table 1: Comparison of datasets: SayPlan Office is the largest environment, BEHAVIOR-1K offers the most long-horizon tasks, and VirtualHome features numerous objects in compact environments.

Data	Tasks	Rooms	Nodes	Actions
SayPlan Office	25	37	202.6	2.1
Behaviour-1k	186	1.23	12.1	4.9
VirtualHome	347	4	195.7	1.6

The SayPlan Office and BEHAVIOR-1K datasets assess general planning capabilities without scene graph augmentation during execution. VirtualHome RobotHow evaluates performance when initial scene graphs lack objects later observed by the agent. Dataset characteristics are summarized in Table 1.

We constructed initial and goal graphs for each task, representing environment states before and after execution. This approach enables automatic validation across large, complex environments, reducing reliance on human evaluation. Combined, the datasets cover 10 environments and 558 tasks paired with initial and goal graphs, highlighting various aspects of embodied planning. Full dataset details are provided in supplementary materials.

SayPlan Office. For evaluating our method on diverse, human-formulated tasks, we used the SayPlan Office Dataset. As the original dataset is unavailable, we reconstructed environment graph representations based on details from the original paper. To align the graph format with our 3DSG structure, we removed pose nodes and directly connected rooms to the scene node. Graph representations and corresponding action sequences were manually constructed, selecting tasks from both simple and complex planning sections of the paper. Using these reconstructed initial graphs and action sequences, we employed a Scene Graph Simulator to generate goal graph representations after task execution. Ambiguous tasks, such as “Put an object into a place where I can enjoy it,” were excluded. The final dataset includes 25 curated tasks with initial and goal graph pairs, detailed in Appendix C.

BEHAVIOR-1K. The Behaviour-1k dataset includes descriptions of 1,000 tasks relevant to real-world scenarios, paired with a simulator providing rooms, scenes, and PDDL task descriptions. We construct graph representations of the environment from PDDL descriptions. The initial graph is derived using a rule-based approach with ontop and inroom predicates. The goal state is generated by GPT-4o (Achiam et al., 2023), which modifies the initial graph based on the task goal and provides human-like task instructions and step-by-step plans. To focus on manipulation tasks, we filter the Behaviour-1k dataset to exclude tasks requiring cooking or cleaning skills, selecting only tasks with predicates `ontop`, `real`, `inside`, `open`, and `toggled on`. This results in 186 tasks with initial and goal graph pairs. The graphs are constructed solely from task descriptions, excluding nodes unrelated to instruction following. This allows for a clear evaluation of planning performance without the need to filter out irrelevant nodes.

VirtualHome RobotHow. For evaluating vision capabilities, we use the VirtualHome simulator, an interactive platform simulating complex household activities. VirtualHome enables interactions such as picking up objects, toggling appliances, and opening appliances, allowing the agent to capture environment images for graph augmentation. We use the RobotHow dataset (Liao et al., 2019), designed for VirtualHome, which includes 1,800 tasks with initial and goal state graphs across 7 home environments. Tasks are filtered to focus on robot-performable manipulation actions, excluding irrelevant tasks (e.g., “Play video game”, “Get shower”) and those involving doors, as they are not represented in the graph structure. Scene descriptions in VirtualHome are translated into graphs using a rule-based approach. Non-grabbable objects are treated as asset nodes, and redundant nodes (e.g., multiple floors, ceilings, walls) are removed for simplicity. The final dataset includes 347 tasks, with duplicates across environments noted. Full details on actions and filtered tasks are in Appendix C.

6 Experiments

6.1 Baselines

We evaluate LookPlanGraph against two baseline methods that incorporate large language models (LLMs) as graph planners.

SayPlan (Rana et al., 2023), operates in two distinct stages: semantic search and iterative replanning. During the semantic search stage, the LLM identifies a minimal sufficient scene graph by expanding room nodes containing relevant items. In the iterative replanning stage, the graph is used to query the same LLM for generating a high-level plan, which is revised based on graph simulation feedback. Notably, both stages are executed sequentially using the same LLM dialogue, with the same prompt being called for both operations.

Since the open-source implementation of SayPlan is unavailable, we developed our own version, referred to as SayPlan*. This version is adapted to process 3DSGs without pose nodes. Additionally, as the graph representation in our context does not require the access functions used for real robots in the original paper, we replaced the access and release function combination with simpler “put on” and “put in” functions.

To reduce the complexity of the task for the LLM and enhance efficiency, we introduced a simplified variant of SayPlan, named **SayPlan Lite**. This approach decomposes the planning process into two distinct LLM dialogues: one dedicated to semantic search and the other to iterative replanning. By separating these tasks and few-shot learning examples, we streamline the planning for LLM, making it more manageable and effective, especially for smaller models. Additionally, we refined the representation of the graph within the prompts to further simplify communication with the LLM. A detailed explanation of these modifications can be found in Appendix A.

6.2 Experimental setup

Our experimental framework rigorously evaluates model performance in generating plans across a diverse set of environments depicted within our dataset. Each task within the experiment framework consists of an input method, an initial environment graph, and corresponding environmental functions. The models engage in iterative processing of these inputs to propose plans or determine actions, which are then executed within a scene graph simulator. The resultant graph is subsequently compared to a predefined goal graph, enabling the computation of various performance metrics (1–4).

To ensure equitable testing conditions for the planning capabilities of different methods, asset exploration is simulated by incorporating objects connected to the asset under investigation. The replanning process is uniformly constrained to a maximum of five iterations across all methods.

The ability to augment graphs was assessed using environments derived from the VirtualHome framework. The scenarios simulated consist of an agent entering a room, initially surveying the surroundings, and subsequently inspecting specific assets. The final node count is compared against the ground truth as obtained from the graph representation. Further details on the VLM prompt structure are provided in Appendix B.

In these experiments, models were utilized, including Llama3.3 (Dubey et al., 2024) with 70 billion parameters, Gemma2 (Team et al., 2024) with 27 billion parameters, and gpt-4o-2024-08-06 (Achiam et al., 2023) for planning tasks. For visual language tasks gpt-4o-2024-08-06 and Llava (Liu et al., 2024a) with 34 billion parameters were employed.

Local models was running on a server equipped with two Tesla V100 GPUs, each with 32GB of VRAM, while other model experiments were conducted via the OpenAI API.

6.3 Metrics

To evaluate the performance of methods based on scene graph we use following metrics. Equations for metric listed in Appendix D.

The **Success Rate** (1) quantifies the percentage of tasks where the method successfully transitions the graph from its initial state to the goal state. A task is considered successful if all nodes are correctly transformed to match their goal configuration.

Average Plan Accuracy (2) evaluates the proportion of correctly modified nodes in the generated plan relative to the total number of modified nodes. This metric measures the precision of the method in altering graph nodes to achieve the goal state.

Average Plan Length (3) reflects the average number of actions required to achieve the goal state across tasks. This metric evaluates the efficiency of the generated plans, with shorter plans generally being preferred, provided they achieve task success.

Node Relevance Ratio (4) measures the method's ability to focus on task-relevant nodes during the planning process. It quantifies the ratio of observed nodes to the number of important nodes, where important nodes are those that differ between the initial and goal graphs. A lower ratio indicates that the method observes fewer unnecessary nodes, which helps reduce token usage for models and improves computational efficiency.

7 Results

The results are divided into three experiments, each addressing a different aspect of the evaluation. Experiment 1 analyses performance across different methods on tasks with static graph where no changes in environment. Experiment 2 analyses the performance of different methods with smaller language models, highlighting their sensitivity to model size and its impact on task success.

7.1 Planning capability across datasets

Table 2 representing performance across datasets. For Llama3.3-70b, SayPlan Lite achieves a success rate of 0.56 and an average plan length of 11.56 in BEHAVIOR-1K but struggles in SayPlan Office with a 0.16 success rate. This suggests methods like SayPlan Lite adapt better to structured tasks but face challenges in contexts requiring nuanced reasoning. Deviations from the original paper's reported 80% success rates likely stem from differences in evaluation metrics, as the original study included human evaluation, which may have considered ambiguous solutions as successful.

Table 2: Methods comparison for Llama3.3.

Method	**SR**	**APA**	**APL**	**NRR**
SayPlan Office				
SayPlan*	0.00	0.00	0.00	-
SayPlan Lite	**0.16**	**0.39**	**4.04**	19.83
LookPlanGraph	0.12	0.34	6.01	**15.72**
BEHAVIOR-1K				
SayPlan*	0.00	0.00	0.00	-
SayPlan Lite	**0.56**	**0.65**	11.56	3.04
LookPlanGraph	0.33	0.41	**10.53**	**2.67**
RobotHow				
SayPlan*	0.00	0.00	0.00	-
SayPlan Lite	**0.39**	**0.41**	**2.08**	40.51
LookPlanGraph	0.25	0.26	3.14	**2.67**

The performance drop in smaller models, such as Gemma 27b, underscores their limitations in managing complex graph-based planning tasks, while larger models like GPT-4o handle multi-step reasoning more effectively. SayPlan Lite's relatively strong performance with smaller models suggests that reducing prompt complexity can partially address model limitations. Dataset characteristics also influence outcomes, with structured tasks in BEHAVIOR-1K yielding better results than the more ambiguous SayPlan Office tasks.

7.2 Impact of LLM on planning performance

Dependence of plan accuracy for different LLM model presented in Table 3. Results highlight the dependency of planning methods on the underlying LLM's size and capabilities. SayPlan* performs well with GPT-4o, achieving 0.48 accuracy, but fails entirely with smaller models like Llama3.3-70b and Gemma 27b. This decline reflects the susceptibility of smaller models to hallucinations, with unreliable function calls and node predictions. While GPT-4o also exhibits some hallucinations, these occur less frequently due to its larger parameter space and improved contextual understanding.

Table 3: Average Plan Accuracy for different models on the SayPlan Office dataset.

Method	**GPT-4o**	**Llama3.3**	**Gemma2**
SayPlan*	0.48	0.00	0.00
SayPlan Lite	0.61	**0.39**	0.00
LookPlanGraph	**0.63**	0.34	**0.15**

SayPlan Lite shows balanced performance, particularly with Llama3.3-70b, where it achieves 0.39 accuracy. Its modularized prompt structure mitigates planning inaccuracies in smaller models. However, its success rate remains low, as shown by its 0.16 score in SayPlan Office. LookPlanGraph demonstrates slightly better accuracy, reaching 0.63 on GPT-4o and 0.33 on Llama3.3-70b.

7.3 Graph augmentation capability

Results for graph augmentation pipeline represented at Table 4. LookPlanGraph$_{gpt-4o}$ demonstrates a relatively modest performance in graph augmentation tasks, with varying levels of success across different relationship types.

Table 4: VirtualHome results on the graph augmentation task. The Node Presence (**NP**) column represents the percentage of nodes from the ground truth number of nodes. The 'On' and 'Inside' columns denote the percentage of nodes with specific relationships.

Method	**NP**	**On**	**Inside**
LookPlanGraph$_{gpt-4o}$	0.66	0.30	0.36
LookPlanGraph$_{llava}$	-	-	-

The LookPlanGraph$_{llava}$ model was able to understand and add nodes in graph augmentation tasks. However, it faces challenges when dealing with larger graphs due to memory constraints. When the number of nodes in the graph becomes too large, the model's performance degrades, and it may stop functioning effectively. This indicates that while the model is good at interpreting tasks, it requires better memory management and scalability optimizations to handle more complex graphs with a higher node count.

Development of SayPlan Lite: We present SayPlan Lite, a streamlined version of the original SayPlan method, designed to boost the efficiency of smaller LLMs for local machine use. Its success showcases the potential for broader application in resource-limited contexts, making it a viable tool for building LLM agents tailored to constrained environments.

Proposal of LookPlanGraph for Dynamic Environments: We propose LookPlanGraph, a graph-based planning framework for dynamic environments. Unlike static scene graphs, this approach initializes with immobile receptacles and dynamically updates with movable objects using a Visual Language Model (VLM) and the agent's egocentric camera. It employs a Memory Graph Mechanism to adapt to environmental changes by focusing on relevant, nearby objects, reducing computational demand. A Graph Augmentation Mechanism further allows real-time exploration and updates, ensuring adaptability to the agent's surroundings.

8 LIMITATIONS

A fundamental aspect of LookPlanGraph is constructing a graph representation of the scene using a 3D Scene Graph structure. This graph-based approach organizes spatial relationships and object states but limits applicability to environments that fit this model. For example, the current representation mainly supports spatial relations like "inside" and "on top", which may not capture the full range of relationships in diverse datasets or real-world scenarios.

LookPlanGraph faces LLM and VLM limitations, including biases, inaccuracies, and incomplete visual inputs. These affect decision-making, especially in complex tasks. Future improvements like fine-tuning, better visual models, or alternative sensory inputs could enhance reliability.

LookPlanGraph assumes perfect low-level action policies, which remains a challenge in robotics. While this simplifies high-level planning, it ignores execution errors, sensor noise, and real-world dynamics. Addressing these challenges would require robust error recovery mechanisms and adaptive control strategies to bridge the gap between high-level plans and real-world execution.

Finally, the Scene Graph Simulator's feedback quality may decline as task complexity grows, particularly with diverse actions and predicates. Developing a more advanced feedback system with improved error detection and corrective dialogue presents a valuable future direction.

9 CONCLUSION

This paper addresses the limitations of static graph representations in dynamic environments by introducing LookPlanGraph, which integrates LLMs and VLMs to update scene graphs in real-time. This approach improves efficiency by focusing on the agent's immediate surroundings while maintaining scalability for smaller LLMs.

Additionally, we present SayPlan Lite, a streamlined version of SayPlan that enhances task decomposition for resource-constrained settings, enabling local execution. Our experiments on SayPlan Office, VirtualHome RobotHow, and BEHAVIOR-1K validate the effectiveness of both methods, demonstrating improved adaptability and efficiency in dynamic environments. These advancements bring LLM-based planning closer to real-world deployment.

10 ETHICAL CONSIDERATIONS

Our approach is based on a large language model that operates in generation mode, and despite the use of a prompt that limits the output format, the model can potentially generate inappropriate and/or offensive output. In addition, language models are prone to hallucinations and can generally produce unforeseen results, so giving them control over mechanisms that could potentially cause harm and testing such mechanisms should be done in a regulated manner, in a specially designated area with limited access to the people involved in the experiments. It is also potentially possible to deliberately execute harmful plans on a robot with the intent to cause harm.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Jae-Woo Choi, Youngwoo Yoon, Hyobin Ong, Jaehong Kim, and Minsu Jang. Lota-bench: Benchmarking language-oriented task planners for embodied agents. *arXiv preprint arXiv:2402.08178*, 2024.

Zhirui Dai, Arash Asgharivaskasi, Thai Duong, Shusen Lin, Maria-Elizabeth Tzes, George Pappas, and Nikolay Atanasov. Optimal scene graph planning with large language model guidance. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14062–14069. IEEE, 2024.

Venkata Naren Devarakonda, Raktim Gautam Goswami, Ali Umut Kaypak, Naman Patel, Rooholla Khorrambakht, Prashanth Krishnamurthy, and Farshad Khorrami. Orionnav: Online planning for robot autonomy with context-aware llm and open-vocabulary semantic scene graphs, 2024. URL `https://arxiv.org/abs/2410.06239`.

Patrick Doherty and Jonas Kvarnstram. Talplanner: A temporal logic-based planner. *AI Magazine*, 22(3):95–95, 2001.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Daniel Ekpo, Mara Levy, Saksham Suri, Chuong Huynh, and Abhinav Shrivastava. Verigraph: Scene graphs for execution verifiable robot planning. *arXiv preprint arXiv:2411.10446*, 2024.

Maria Fox and Derek Long. Pddl2. 1: An extension to pddl for expressing temporal planning domains. *Journal of artificial intelligence research*, 20:61–124, 2003.

Paul Gay, James Stuart, and Alessio Del Bue. Visual graphs from motion (vgfm): Scene understanding with object geometry reasoning. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pp. 330–346. Springer, 2019.

Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning, 2023. URL `https://arxiv.org/abs/2309.16650`.

Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. *arXiv preprint arXiv:2311.17842*, 2023.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, pp. 9118–9147. PMLR, 2022.

Ue-Hwan Kim, Jin-Man Park, Taek-Jin Song, and Jong-Hwan Kim. 3-d scene graph: A sparse and semantic representation of physical environments for intelligent agents. *IEEE transactions on cybernetics*, 50(12):4921–4933, 2019.

Andrey Kurenkov, Roberto Martín-Martín, Jeff Ichnowski, Ken Goldberg, and Silvio Savarese. Semantic and geometric modeling with neural message passing in 3d scene graphs for hierarchical mechanical search. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11227–11233. IEEE, 2021.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, et al. Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation. *arXiv preprint arXiv:2403.09227*, 2024.

Yuan-Hong Liao, Xavier Puig, Marko Boben, Antonio Torralba, and Sanja Fidler. Synthesizing environment-aware activities via activity sketches. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6284–6292, 2019. doi: 10.1109/CVPR.2019.00645.

Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. Llm+p: Empowering large language models with optimal planning proficiency, 2023. URL `https://arxiv.org/abs/2304.11477`.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a.

Hengyue Liu, Ning Yan, Masood Mortazavi, and Bir Bhanu. Fully convolutional scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11546–11556, 2021.

Yuchen Liu, Luigi Palmieri, Sebastian Koch, Ilche Georgievski, and Marco Aiello. Delta: Decomposed efficient long-term robot task planning using large language models, 2024b. URL `https://arxiv.org/abs/2404.03275`.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8494–8502, 2018.

Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian D Reid, and Niko Suenderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable task planning. *CoRR*, 2023.

Antoni Rosinol, Andrew Violette, Marcus Abate, Nathan Hughes, Yun Chang, Jingnan Shi, Arjun Gupta, and Luca Carlone. Kimera: From slam to spatial perception with 3d dynamic scene graphs. *The International Journal of Robotics Research*, 40(12-14):1510–1546, 2021.

Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11523–11530. IEEE, 2023.

Ying Su, Zhan Ling, Haochen Shi, Jiayang Cheng, Yauwai Yim, and Yangqiu Song. Actplan-1k: Benchmarking the procedural planning ability of visual language models in household activities. *arXiv preprint arXiv:2410.03907*, 2024.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Zixuan Wang, Bo Yu, Junzhe Zhao, Wenhao Sun, Sai Hou, Shuai Liang, Xing Hu, Yinhe Han, and Yiming Gan. Karma: Augmenting embodied ai agents with long-and-short term memory systems. *arXiv preprint arXiv:2409.14908*, 2024.

A APPENDIX – PROMPT STRUCTURES

A.1 LOOKPLANGRAPH

A.1.1 STATIC PROMPT

The static prompt remains constant across all tasks and provides foundational information to the LLM. It includes the agent's role and objectives, a description of states and relationships that can appear in the JSON graph representation, a list of functions available to the agent (e.g., ”look on,” ”look inside,” ”pick up”), the expected output format (structured JSON response detailing the next action), and two examples of how the agent should respond.

After the static prompt, dynamic components follow. These include the instruction, which is a natural language description of the task, a filtered JSON graph representation, and feedback. The JSON graph is simplified to include only the nodes and attributes relevant to performing the action, such as those in the same room as the agent or objects the agent interacted with earlier in the task. This filtering ensures the prompt remains concise while providing necessary context for long-horizon tasks.

Agent Role: You are an expert in graph-based task planning. Given a graph representation of the environment, your goal is to generate a next move for the agent to follow to solve the given instruction.

Graph environment states:

- ontop_of(<asset>): Object is located on <asset>.
- inside_of(<asset>): Object is located inside <asset>.
- closed: Asset can be opened.
- open: Asset can be closed or kept open.
- on: Asset is currently on.
- off: Asset is currently off.

Available Functions:

- go_to(<room>): Move the agent to room node. Use it only with room nodes.
- pick_up(<object>): Pick up an accessible object from the accessed node. You can handle only one item.
- put_on(<asset>): Put held object on <asset>.
- put_inside(<asset>): Put held object inside of <asset>.
- turn_on/off(<node>): Toggle object on or off.
- open/close(<node>): Open or close node.
- look_on(<asset>): Look on top of <asset>. Adds the discovered objects to the memory graph.
- look_inside(<asset>): Look inside of <asset>. Adds the discovered objects to the memory graph.
- done(<node>): Call this function with any node when the goal has been achieved.

Answer only with JSON without comments. Output Response Format:

```
{
  "chain_of_thought": Break down your reasoning into intermediate steps.
  "next_move": {
    "function_name": Name of the function from Available Functions.
    "function_target": Node name.
  }
}
```

Examples of output:

```
{
  "chain_of_thought": [
    "i have found the coffee mug,
    the coffee machine and tom's wardrobe on the graph",
    "collect coffee mug",
    "generate plan for making coffee",
    "place coffee mug on Tom's wardrobe"
  ],
  "next_move": {
    "function_name": "go_to",
    "function_target": "bobs_room"
  }
```

```
}

{
  "chain_of_thought": [
    "goal is reached",
    "i am inside bobs_room",
    "now i call function to show thats i am done with task"
  ],
  "next_move": {
    "function_name": "done",
    "function_target": "bobs_room"
  }
}
```

A.1.2 Dynamic prompt example

Instruction: Take the socks, bottle of perfume, toothbrush, and notebook out of the carton and place them on the sofa in the living room.

Memory graph:

```
{"nodes":{"room":[
{"id":"living_room1"}],
"asset":[{"id":"floor1","located":"living_room1","states":[]},
{"id":"sofa1","located":"living_room1","states":[]}],"object":[
{"id":"carton1","relation":"ontop_of","related_to":"sofa1","states":["closed"]},
{"id":"sock1","relation":"ontop_of","related_to":"sofa1","states":[]},
{"id":"sock2","relation":"ontop_of",
"related_to":"sofa1","states":[]},
{"id":"bottle__of__perfume1","relation":"ontop_of",
"related_to":"sofa1","states":["closed"]},
{"id":"toothbrush1","relation":"ontop_of","related_to":"sofa1","states":[]},
{"id":"notebook1","relation":"ontop_of","related_to":"sofa1","states":[]}],
"agent":[{"id":"agent1","location":"living_room1","holding":""}]}}
```

A.2 SayPlan Lite

SayPlan Lite splits the prompt into two stages corresponding to SayPlan's workflow, hiding irrelevant information at each stage and separating the LLM's API interactions into two parts. This approach minimizes potential hallucinations.

A.2.1 Semantic search

```
Agent Role:
You are an efficient graph search agent tasked with exploring
a graph-based environment to  find specific items based on a given instruction.
You interact with the environment via an API to expand or contract room nodes.
Objective:
Your goal is to identify the relevant
parts of the graph to fulfill the instruction.
You must expand appropriate room nodes, filter out irrelevant ones,
and verify the graph using the environment's API.

Environment API:
expand_node(<room>): Reveal assets/objects connected to a room node.
contract_node(<room>): Hide assets/objects, reducing
graph size for memory constraints.
verify_plan(): Verify graph in the scene graph environment.

Guidelines:
```

```
1. Do not expand asset or object nodes, only room nodes.
2. Contract irrelevant nodes to reduce memory usage.
3. Once all relevant objects are found, use verify_plan() to confirm that graph
is rellevant to the task.

Output Response Format: Your response should follow this structure:
{
"chain_of_thought": break your problem down into a series of intermediate
reasoning steps to help you determine your next command,
"reasoning": justify why the next action is important
"command":
  {
  "command_name": Environment API call
  "node_name": node to perform an operation on
  }
}

Example of output:
{
  "chain_of_thought": [
    "i have found a wardrobe in tom’s room",
    "leave this node expanded",
    "the coffee mug is not in his room",
    "still have not found the coffee machine",
    "kitchen might have coffee machine and coffee mug",
    "explore this node next"
  ],
  "reasoning": "i will expand the kitchen next",
  "command": {
    "command_name": "expand_node",
    "node_name": "kitchen1"
  }
}
```

A.2.2 ITERATIVE RE-PLANNING

```
Agent Role: You are an expert in graph-based task planning.
Given a graph representation of the environment,
your goal is to generate a precise, step-by-step task plan
for the agent to follow and solve the given instruction.

Graph environment states:
ontop_of(<asset>): Object is located on <asset>
inside_of(<asset>): Object is located inside <asset>
attached_to(<asset>): Object is attached to <asset>
closed: Asset can be opened
open: Asset can be closed or kept open
on: Asset is currently on
off: Asset is currently off

Available Functions (use these exclusively for planning):
go_to(<room>): Move the agent to room node. Use it only with room nodes.
pick_up(<object>): Pick up an accessible object from the accessed node.
You can handle only one item.
put_on(<asset>): Put holded object on asset.
put_inside(<asset>): Put holded object inside of asset.
put_under(<asset>): Put holded object under of asset.
attach(<asset>): Attach holded object to asset.
```

```
turn_on/off(<object>): Toggle object at agent’s node,
if accessible and has affordance.
open/close(<node>): Open/close node at agent’s node, affecting object.

Answer only with JSON without comments. Output Response Format:
{"chain_of_thought": Break down your reasoning into intermediate steps.
"plan": List the environment function calls to solve the task.}

Example of output:
{
  "chain-of-thought": [
    "i have found the coffee mug,
    the coffee machine and tom’s wardrobe on the graph",
    "collect coffee mug",
    "generate plan for making coffee",
    "place coffee mug on Tom’s wardrobe"
  ],
  "plan": [
    "go_to(bobs_room1)",
    "pick_up(coffee_mug1)",
    "go_to(kitchen1)",
    "put_inside(coffee_machine1)",
    "turn_on(coffee_machine1)",
    "turn_off(coffee_machine1)",
    "pick_up(coffee_mug1)",
    "go_to(toms_room1)",
    "put_on(wardrobe2)"
  ]
}
```

B APPENDIX – VLM PROMPT STRUCTURE

B.1 PROMPT

Describe the image.

```
Return the results in a predefined JSON format as follows:
[
  {
    "name": "object_name",
    "relation": "relation_type",
    "related_to": "related_object_name",
    "states": "object_state",
    "properties": "object_properties"
  }
]
```

Guideline:

```
1. Include only objects that can be moved.
2. Possible states are: open, closed, turned_on, turned_off.
3. Possible relations are: ontop_of, inside_of.

Guidelane:
1. Include only objects that can be moved.
2. Possible states are: open, closed, turned_on, turned_off.
3. Possible relations are: ontop_of, inside_of.

Example Output:
```

```
[
  {
    "name": ["bowl",1],
    "relation": "ontop_of",
    "related_to": ["bench", 1],
    "states": "",
    "properties": "black"
  },
  {
    "name": ["apple",1],
    "relation": "inside_of",
    "related_to": ["bowl", 1],
    "states": "",
    "properties": "red"
  },
  {
    "name": ["apple",2],
    "relation": "inside_of",
    "related_to": ["bowl",1],
    "states": "",
    "properties": "green"
  },
  {
    "name": ["bottle",1],
    "relation": null,
    "related_to": null,
    "states": "closed",
    "properties": "green"
  }
]
```


```
Do not add objects from list:
<list of assets in the same room and already founded objects>
```

C APPENDIX – DATASET PREPARATION

The graphs is implemented using the NetworkX library, tool for creating and manipulating graph data structures.

Each node in the graph has specific attributes based on its type. These attributes provide detailed information that can be used in the planning and task execution process:

- **Scene Node:** Contains the name of the scene.
- **Place Node:** Contains the scene it belongs to, identifying its broader location in the environment.
- **Asset Node:** Includes the location of the asset, its current state, allowed actions, and any other properties relevant to its use in the environment.
- **Object Node:** Describes the relationship type with other nodes, the asset it belongs to, its states, allowed actions, and properties.
- **Agent Node:** Tracks the location of the agent and the item currently held by the agent.

Filtered actions from VirtualHome:

Open door, Lock door, Look out window, Movie, Clean, Write school paper, Dust, Play games, Get dressed, Playing video game, Shave, Print out papers, Watch fly, Walk through, Admire art, Gaze out window, Look at painting, Check appearance in mirror, Shut front door, Look at mirror, Write an email, Browse internet, Watch TV, Take shower, Work, Drink, Wash teeth, Wash dishes by hand, Pet cat, Brush teeth, Keep an eye on stove as something is cooking, Open front door, Close door, Pick up phone.

Limited actions for VirtualHome:

”FIND”, ”WALK”, ”GRAB”, ”SWITCHON”, ”TURNTO”, ”PUTBACK”, ”LOOKAT”, ”OPEN”, ”CLOSE”, ”PUTOBJBACK”, ”SWITCHOFF”, ”PUTIN”, ”RUN”,

SayPlan Office tasks: SayPlan Office tasks instructions listed in Table 5.

Table 5: SayPlanOffice Tasks

Tasks description
Close Jason’s cabinet.
Refrigerate the orange left on the kitchen bench.
Take care of the dirty plate in the lunchroom.
Place the printed document on Will’s desk.
Peter is working hard at his desk. Get him a healthy snack.
Hide one of Peter’s valuable belongings.
Wipe the dusty admin shelf.
There is coffee dripping on the floor. Stop it.
Place Will’s drone on his desk.
Move the monitor from Jason’s office to Filipe’s.
My parcel just got delivered! Locate it and place it in the appropriate lab.
Check if the coffee machine is working.
Heat up the chicken kebab.
Something is smelling in the kitchen. Dispose of it.
Heat up the noodles in the fridge, and place it somewhere where I can enjoy it.
Throw the rotting fruit in Dimity’s office in the correct bin.
Safely file away the freshly printed document in Will’s office, then place the undergraduate thesis on his desk.
Make Niko a coffee and place the mug on his desk.
Tobi spilt soda on his desk. Throw away the can and take him something to clean with.
I want to make a sandwich. Place all the ingredients on the lunch table.
Empty the dishwasher. Place all items in their correct locations.
A delegation of project partners is arriving soon. We want to serve them snacks and non-alcoholic drinks. Prepare everything in the largest meeting room. Use items found in the supplies room only.
Serve bottled water to the attendees who are seated in meeting room 1. Each attendee can only receive a single bottle of water.
Locate all 6 complimentary t-shirts given to the PhD students and place them on the shelf in admin.
I’m at the lunch table. Let’s play a prank on Niko. Dimity might have something.

D APPENDIX – METRICS EQUATIONS

$$\text{SR} = \frac{\text{Number of successful tasks}}{\text{Number of tasks}}, \tag{1}$$

$$\text{APA} = \frac{\text{Number of right changed nodes}}{\text{Number of changed nodes}}, \tag{2}$$

$$\text{APL} = \frac{\sum \text{Plan length}}{\text{Number of tasks}}, \tag{3}$$

$$\text{NRR} = \frac{\text{Observed nodes}}{\text{Important nodes}}. \tag{4}$$