

# DKPROMPT: Domain Knowledge Prompting Vision-Language Models for Open-World Planning

Anonymous Author(s)

Affiliation

Address

email

**Abstract:** Vision-language models (VLMs) have been applied to robot task planning problems, where the robot receives a task in natural language and generates plans based on visual inputs. While current VLMs have demonstrated strong vision-language understanding capabilities, their performance is still far from being satisfactory in planning tasks. At the same time, although classical task planners, such as PDDL-based, are strong in planning for long-horizon tasks, they do not work well in open worlds where unforeseen situations are common. In this paper, we propose a novel task planning and execution framework, called DKPROMPT, which automates VLM prompting using domain knowledge in PDDL for classical planning in open worlds. Results from quantitative experiments show that DKPROMPT outperforms classical planning, pure VLM-based and a few other competitive baselines in task completion rate.<sup>1</sup>

**Keywords:** AI Planning, Vision-language Models, Open World

## 1 Introduction

Prompting foundation models such as large language models (LLMs) and vision-language models (VLMs) requires extensive domain knowledge and manual efforts, resulting in the so-called “prompt engineering” problem. To improve the performance of foundation models, one can provide examples explicitly [1] or implicitly [2], or encourage intermediate reasoning steps [3, 4]. Despite all the efforts, their performance in long-horizon reasoning tasks is still limited. Classical planning methods, including those defined by Planning Domain Definition Language (PDDL), are strong in ensuring the soundness, completeness and efficiency in planning tasks [5]. However, those classical planners rely on predefined states and actions, and do not perform well in open-world scenarios. We aim to enjoy the openness of VLMs in scene understanding while retaining the strong long-horizon reasoning capabilities of classical planners. Our key idea is to extract domain knowledge from classical planners for prompting VLMs towards enabling classical planners that are visually grounded and responsive to open-world situations.

Given the natural connection between planning symbols and human language, this paper investigates how pre-trained VLMs can assist the robot in realizing symbolic plans generated by classical planners, while avoiding the engineering efforts of checking the outcomes of each action. Specifically, we propose a novel task planning and execution framework called DKPROMPT, which prompts VLMs using domain knowledge in PDDL, generating visually grounded task planning and situation handling. DKPROMPT leverages VLMs to detect action failures and verify action affordances towards successful plan execution. We take the advantage of the domain knowledge encoded in classical planners, including the actions defined by their effects and preconditions. By simply querying current observations against the action knowledge, similar to applying VLMs to Visual Question

---

<sup>1</sup><https://dkprompt.github.io/>

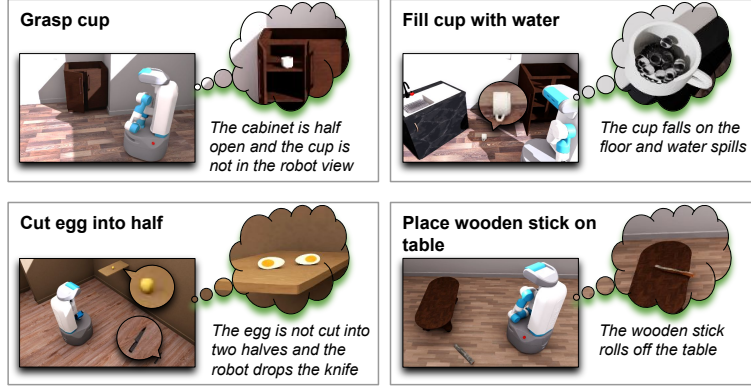


Figure 1: A few unforeseen situations during action execution. In the top-left example, the robot “opened” the cabinet door to get prepared for grasping the cup. It was expected that the cup in white would have been in the robot’s view after the “opening” action, while a situation occurred, i.e., the cabinet was only half-open. DKPROMPT prompts vision-language models (VLMs) using domain knowledge to detect and address such situations. While one can develop a safeguard to detect cabinet opening being successful, our goal is to automate this process, avoiding such manual efforts and handling unforeseen situations.

36 Answering (VQA) tasks, DKPROMPT can trigger the robot to address novel situations and recover  
37 from action failures.

38 We conducted quantitative evaluations using the OmniGibson simulator [6]. We assume that robot  
39 actions are *imperfect* by nature, frequently causing *situations*<sup>2</sup> during execution (Figure 1). Results  
40 demonstrate that DKPROMPT utilizes domain knowledge to adaptively generate task plans, recovers  
41 from action failures and re-plans when situations occur. In addition, we hope that researchers  
42 working on VLMs, robot planning or both find our evaluation platform useful for their research. In  
43 particular, the open-world situations and structured world knowledge presents a new playground for  
44 comparing robot planning and vision-language understanding using large-scale models.

## 45 2 Related Work

46 This section starts with covering a wide range of downstream applications of classical AI planners in  
47 symbolic task planning. It then explores the role of Large Language Models (LLMs) in robot plan-  
48 ning, discussing their strengths (e.g., rich in commonsense) and limitations (e.g., lack of correctness  
49 guarantee). Finally, it examines the recent advancements in Vision-language Models (VLMs) and  
50 their impact on the robotics community.

### 51 2.1 Classical AI Planning for Robots

52 Automated planning algorithms have a long-standing history in the literature of symbolic AI and  
53 have been widely used in robot systems. Shakey is the first robot that was equipped with a planning  
54 component, which was constructed using STRIPS [8]. Recent classical planning systems designed  
55 for robotics commonly employ Planning Domain Description Language (PDDL) or Answer Set  
56 Programming (ASP) as the underlying action language for planners [9, 10, 11, 12, 13, 14, 15, 16,  
57 17, 18]. Most classical planning algorithms that are designed for robot planning do not consider  
58 perception. Though some recent works have already shown that training vision-based models from  
59 robot sensory data can be effective in plan feasibility evaluation [19, 20, 21, 22, 23], their methods  
60 did not tightly bond with language symbols which are the state representations for classical planning  
61 systems. The most relevant work to our study is probably the research by Migimatsu and Bohg,  
62 which trained domain-specific predicate classifiers from webscale data and deployed on a robot  
63 planning system [24]. We propose DKPROMPT that investigates how off-the-shelf Vision-language  
64 Models connect perception with symbolic language which is used to represent robot knowledge.

<sup>2</sup>Situation is an unforeseen world state that potentially prevents an agent from completing a task using a solution that normally works [7].

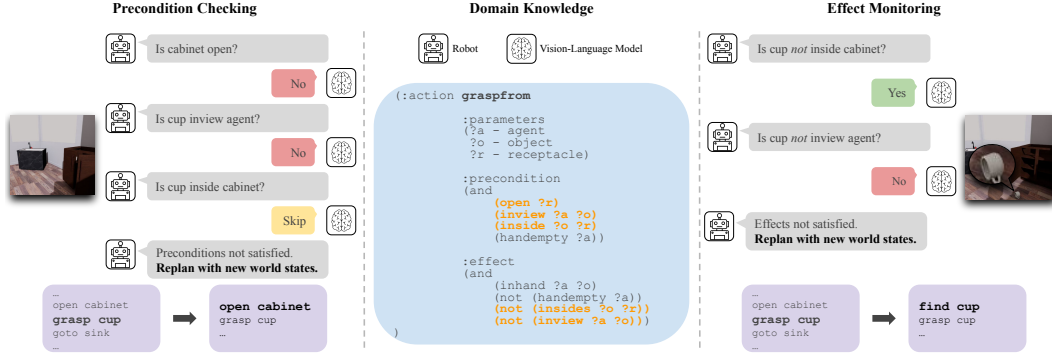


Figure 2: An overview of DKPROMPT. By simply querying the robot’s current observation against the domain knowledge (i.e., action preconditions and effects) as VQA tasks, DKPROMPT can call the classical planner to generate a new valid plan using updated world states. Note that DKPROMPT only queries about predicates. The left shows how DKPROMPT checks every precondition of the action to be executed next, and the right shows how it verifies the expected action effects are all in place after action execution. Replanning is triggered when preconditions or effects are unsatisfied after updating the planner’s action knowledge.

## 2.2 Classical Planning with Large Language Models

In the light of the recent advancement in artificial intelligence, many LLMs have been developed in recent years [25, 26, 27, 28]. These LLMs can encode a large amount of common sense [29] and have been widely applied to robot task planning [30, 31, 32, 33, 34, 35, 36, 37, 38]. However, a major drawback of existing LLMs is their lack of long-horizon reasoning/planning abilities for complex tasks [39, 40, 41]. Specifically, the output they produce when presented with such a task is often incorrect in the sense that following the output plan will not actually solve the task. As a result, a wide range of studies have investigated approaches that combine the classical planning methodology with LLMs in robotic domains [42, 43, 44, 45, 46, 47, 48, 49, 50, 51]. However, neither LLMs nor classical planners are inherently *grounded*, often necessitating complex interfaces to bridge the symbolic-continuous gap between language and robot perception. Our approach seeks to ground classical planners by utilizing pre-trained VLMs through a novel but straightforward domain knowledge prompting strategy.

## 2.3 Vision-language Models in Robotics

VLMs have emerged as powerful methods that integrate visual and linguistic information for complex AI tasks [52, 53, 54, 55, 56]. Researchers have started to employ such models in robot systems [57, 58, 59, 60, 61], where these models have shown effectiveness in, for example, semantic scene understanding [62], open-ended agent learning [63], guiding robot navigation [64] and manipulation behaviors [65, 66]. Recent VLMs have also been used for building *planning* frameworks [67, 68]. Adaptive planning significantly improve task performance through better environment awareness and fault recovery, and language understanding allows robots to seek human assistance in handling uncertainty [69, 70]. There have been recent methods, similar to us, that query VLMs for action success, failures, and affordances [71, 72, 73]. Different from their work, DKPROMPT uses classical planners to generate executable symbolic plans rather than solely relying on pre-trained models. Additionally, DKPROMPT integrates domain knowledge into the prompts, enhancing the grounded connection between the VLMs and the symbolic planner.

## 3 DKPROMPT for Planning in Open Worlds

This section presents the implementation details of DKPROMPT in a robot planning system, particularly suitable for open worlds. The system assumes that the agent is equipped with a predefined set of actions which are *imperfect* by nature, frequently causing unforeseen situations (Section 3.1). The agent also processes a handful of *action knowledge*, where actions are defined by their preconditions and effects. These preconditions and effects are further represented as *objects* and *propositions*, i.e.,

Table 1: Actions, constraints, and their uncertain outcomes (i.e., situations).

Actions	Constraints	Situations
find	(1) The object and the agent are in the same room.	(1) The robot succeeds in navigation but the object is not inview. (2) There is no free space near the object so navigation fails. (3) The object that the robot is holding drops during navigation.
grasp	(1) The object is inview. (2) The agent’s hand is empty.	(1) The robot fails to grasp, and the object position remains unchanged. (2) The robot fails to grasp, and the object drops nearby.
placein	(1) The object is inhand. (2) The receptacle is inview. (3) The receptacle is not closed.	(1) The robot fails to place, and the object remains in the robot’s hand. (2) The robot fails to place, and the object drops nearby.
placeon	(1) The object is inhand. (2) The receptacle is inview.	(1) The robot fails to place, and the object remains in the robot’s hand. (2) The robot fails to place, and the object drops nearby.
fillsink	(1) The sink is inview.	(1) The robot fails to open the faucet.
fill	(1) The container is inhand. (2) The agent is near sink. (3) The container is empty.	(1) The container is not fully filled. (2) The container drops nearby.
open	(1) The object is inview.	(1) The robot fails to open and the object remains closed.
close	(1) The object is inview.	(1) The robot fails to close and the object remains open.
turnon	(1) The object is inview.	(1) The robot fails to turn on the switch and the object remains off.
cut	(1) The object is inview. (2) A knife is inhand.	(1) The object is not cut into half, and the knife is still in the robot’s hand. (2) The object is not cut into half, and the knife drops nearby.

97 predicates (Section 3.2). We then introduce how DKPROMPT takes advantages of the action knowl-  
98 edge for states update and online re-planning (Section 3.3).

### 99 3.1 Robot Actions

100 Our system considers ten actions (as listed in Table 6), including basic navigation and manipulation.  
101 Situations occur after actions are successfully triggered by the agent. Table 6 also provides examples  
102 of situations that happen following specific actions. Some of these situations impact the world  
103 states, while others do not. For example, the robot may fail on a “grasp” action, resulting in the  
104 target object, originally on the table, to fall on the floor nearby (changing the state from `on(obj, table)`  
105 to `on(obj, floor)`). On the other hand, the object might also remain on the table with  
106 the world states being unchanged. To quantify the openness of different environments, we created  
107 the simulation platform in such a way that one can easily adjust the probability of a situation’s  
108 occurrence. The source code of our benchmark system will be made available in our project website.

109 Actions are implemented in a discrete manner for simplification purposes, since continuous action  
110 execution is not this paper’s focus. For instance, “find” action is implemented by teleporting the  
111 agent from its initial position to a randomly-sampled obstacle-free goal position near the target, and  
112 “fill” action is by adding fluid particles directly into the container that the robot is holding.

113 Actions are subject to several constraints. For example, “grasp” action is deemed executable only if  
114 the target object is in the agent’s view (assuming vision-based manipulation) and the agent’s hand is  
115 empty. Similarly, “cut” action is considered executable only if the object to be cut is in the agent’s  
116 view and the agent is currently holding a knife. Calling an action with at least one unsatisfied  
117 constraint will result in an action failure, but without any changes to the world states. Note that such  
118 constraints are not made available to agents, instead, they are partially encoded as domain (action)  
119 knowledge that the agent possesses.

120 We assume that situations can only happen during action execution, but are only observable by  
121 agents either before or after the action execution phase. This assumption indicates that situations  
122 are solely caused by actions, and we are aware of a few recent robotic research that have started  
123 to consider more generalized situation handling [7]. We leave situations that caused by external  
124 environmental factors (human or other embodiments) to future work.

### 125 3.2 Predicates

126 A single action is usually defined by multiple preconditions and effects in the domain knowledge.  
127 VLMs, especially for those that are not trained using domain-specific data, frequently produce in-

Table 2: DKPROMPT assumptions for predicates.

<b>Perceptible in vision</b>	inview, closed, open, inside, halved, onfloor, ontop, cooked
<b>Perceptible in non-vision</b>	handempty, inhand, hot
<b>Imperceptible</b>	turnedon, filled, inroom, insource

accurate answers that cause disagreements among the given preconditions (or effects). For instance, the VLM might answer “Yes” to both `on(apple, table)` and `inhand(apple)` after the robot picks up an apple from the table. In this paper, DKPROMPT categorizes predicates into three: *perceptible in vision*, *perceptible in non-vision*, and *imperceptible*. DKPROMPT will only ask about “*perceptible in vision*” predicates. Intuitively, we believe VLMs should be and will be only good at visually-perceptible predicates. The robot will then have ground truth access to *perceptible in non-vision* predicates (this assumption also applies to all other baselines). We leave identifying these predicates using more advanced Multimodal Language Models to future work. As for the remaining *imperceptible* predicates, the DKPROMPT agent maintains a positive attitude and assumes they are always True. This suggests that DKPROMPT believes these predicates will never be affected by any situation.

### 3.3 DKPROMPT

Before every action execution, DKPROMPT extracts knowledge about action preconditions from the planner’s domain description. For instance, as indicated in Figure 2, action `graspfrom(a, o, r)` has preconditions of `open(r)`, `inview(a, o)`, `inside(o, r)`, and `handempty(a)`, meaning that to grasp an object  $o$  from a receptacle  $r$ ,  $r$  should be open (not closed),  $o$  should be in the agent’s current first person view,  $o$  should be inside  $r$ , and the agent’s hand should be empty. Then, we simply convert each action precondition into a natural language query by using manually defined templates, though it has been evident that LLMs can be used for the translation between PDDL and natural language [48]. Examples include “*Is <math>o</math> inview agent?*” and “*Is <math>o</math> inside <math>r</math>?*” Paring each natural language query with the current observation from the robot’s first-person view, we call the VLM to get answers indicating if the precondition is satisfied.

According to the results (i.e., “yes”, “no”, or “skip” if unsure) from the VLM, DKPROMPT will update the current state information in the classical planning system. Figure 2 (Left) shows an example where the robot wants to `graspfrom(cup, cabinet)` but fails to detect “cabinet is open”, “cup is inview of agent”, and is suspicious about if “cup is in the cabinet” (the VLM answers “skip” to this question) given the current observation. As a result, DKPROMPT will update the current state by changing `open(cabinet)` to `closed(cabinet)`, and removing `inview(agent, cup)`. `inside(cup, cabinet)` will remain the same because we do not update the state if the VLM answers “skip”, indicating the agent holds a positive attitude that situations will not commonly occur. We then provide the updated world state to the classical planner as the “new” initial state to re-generate a plan. In the above example, instead of `graspfrom(cup, cabinet)`, the robot will now take the action of `open(cabinet)` again according to the newly-generated action plan. After every action execution, DKPROMPT extracts knowledge about action effects from the planner’s domain description, illustrated in Figure 2 (Right). It queries action effects by using the VLM. If the effects are not satisfied, the robot will update its belief on the current states and re-plan accordingly. The knowledge-based automated prompting strategy of VLMs enables our planning system to adaptively capture and handle unforeseen situations at execution time.

## 4 Experiments

We conducted extensive experiments to evaluate the performance of DKPROMPT comparing with baselines from the literature. Our hypothesis is DKPROMPT produces the highest task completion rate because of its effectiveness in plan monitoring and online re-planning using domain knowledge and perception.

Table 3: Task descriptions and initial plan length.

Name	Descriptions	Initial plan length
boil water in the microwave	Pick up an empty cup in a closed cabinet, fill it with water using a sink, and boil it in a microwave.	12
bring in empty bottle	Find two empty bottles in the garden and bring them inside.	8
cook a frozen pie	Take an apple pie out of the fridge and heat it using an oven.	8
halve an egg	Find a knife in the kitchen and use it to cut a hard-boiled egg into half.	4
store firewood	Collect two wooden sticks and place them on a table.	8

## 4.1 Experiment Setup

Quantitative evaluation results are collected in the OmniGibson simulator [6]. The agent is equipped with a set of skills, and aims to use its skills to interact with the environment, completing long-horizon tasks autonomously. In the experiment, we consider five everyday tasks that are “boil water in the microwave”, “bring in empty bottle”, “cook a frozen pie”, “halve an egg”, and “store firewood”. Their detailed descriptions are shown in Table 3. These five tasks are originally from the Behavior 1K benchmark [6] that are accompanied with the simulator. Task descriptions including initial and goal states are written in PDDL and symbolic plans are generated using the fast-downward planner [74].

## 4.2 Results

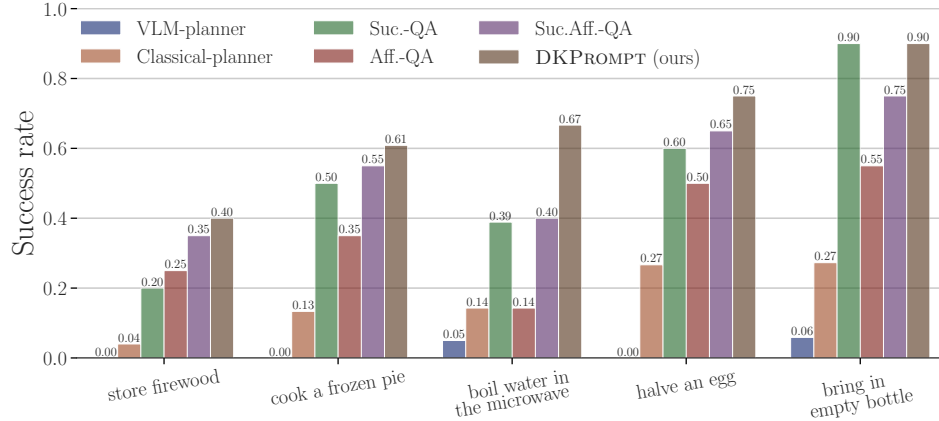


Figure 3: DKPROMPT v.s. baselines in success rate over five everyday tasks.

**Comparisons with Baselines.** Figure 3 presents the main experimental results and details the comparative success rates of various methods from the literature. The methods include:

- VLM-planner, which uses the VLM as a planner to generate task plans, similar to [31]. For fair comparisons, we also provide domain knowledge (as natural language) in the prompts for the VLM.
- Classical-planner, which is a typical classical planning approach without perception, assuming all action executions are successful;
- Suc.-QA, which uses a classical planner to generate plans, and asks about action success after each action execution. This baseline is inspired by [71], and we use the same query provided in their paper, which is “*Did the robot successfully <action>?*” Suc.-QA does not consider if the next action is executable;
- Aff.-QA, which uses a classical planner to generate plans, and asks about action affordance before each action execution. This baseline is designed with prompts provided in the original PaLM-E paper [72], which are “*Is it possible to <action> here?*” and “*Was <action> successful?*” Aff.-QA does not consider whether the previous action is successful;



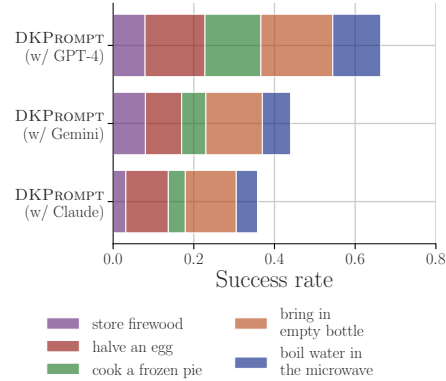
Table 4: Ablation study on preconditions and effects.

#	Methods	Tasks					avg. (%)
		boil water in the microwave	bring in empty bottle	cook a frozen pie	halve an egg	store firewood	
<b>Ours</b>							
1	DKPROMPT	<b>66.7</b>	90.0	<b>60.9</b>	<b>75.0</b>	<b>40.0</b>	<b>66.5</b>
<b>Ablation</b>							
2	Eff.-only	50.0	<b>93.8</b>	26.7	66.7	28.0	53.0
3	Pre.-only	17.6	75.0	35.0	55.0	20.0	41.5

- Suc.Aff.-QA, which uses a classical planner, asks about both action affordance (before each action execution) and action success (after each action execution), similar to [33].<sup>3</sup>

When using VLM itself as the planner, the agent frequently fails in finding an executable plan, resulting in the lowest success rate. This finding is consistent with recent work [39] and motivates the development of other research that combines classical planning with large models [48]. Classical planner, which operates without visual feedback during task execution, shows the second lowest success rate across five tasks compared to other evaluated methods, highlighting its limited effectiveness in handling situations and recovering from potential action failures. In contrast, methods that involve querying for action affordances, success probabilities, or both, achieve much higher success rates as compared to the “blind” classical planning approach. This improvement demonstrates the general advantage of incorporating visual feedback and high-level reasoning in task planning systems. While it is always a good practice to verify both before and after an action (like Suc.Aff.-QA), we found that Suc.-QA also surpasses the performance of Aff.-QA, indicating that there is a greater positive impact on task completion from action failure recovery, and VLMs have better zero-shot reasoning capabilities on the direct effects caused by actions.

We observed that DKPROMPT consistently outperforms baselines in task completion rates, which supports our hypothesis. By incorporating domain knowledge (i.e., action preconditions and effects) for prompting, DKPROMPT is significantly better than other methods, including Suc.Aff.-QA that also cares about affordance prediction and failure detection. However, Suc.Aff.-QA queries about actions solely by their names, which provides less information than the detailed domain knowledge used by DKPROMPT, indicating that action knowledge is more informative for pretrained VLMs to reason over.



**Ablation Study on Preconditions and Effects.** Table 4 presents an ablation study comparing the performance of different versions of our approach across the same set of tasks. DKPROMPT integrates both action effects and action preconditions, while we are also curious to know how they affect the overall task completion independently. DKPROMPT achieves an average success rate of 66.5%. For ablation methods where only action effects are considered (Eff.-only), the average success rate drops to 53.0%, and for methods considering only preconditions (Pre.-only), it further decreases to 41.5%. This suggests that the integration of both effects and preconditions in DKPROMPT significantly enhances task performance compared to considering these components separately.

**Performance of Other VLMs.** We also run experiments on various VLMs, including GPT-4 (as being used in the original implementation of DKPROMPT) from OpenAI [41], Gemini 1.5 from Google [75], and Claude 3 from Anthropic. According to Figure 4, GPT-4 consistently performs better than Gemini and Claude. By looking at the highest accuracy among all the VLMs (i.e., less than 65%), our evaluation benchmark (designed with challenging open-world situations and rich domain knowledge) presents a simulation platform, dataset and success criteria that other researchers

<sup>3</sup>We use the same VLM as ours (GPT4) for implementing all baselines that require a VLM.

working on AI planning, VLMs or both might find useful. We will open source the benchmark including software and data to the public after the anonymous review phase.

### 4.3 Real-Robot Deployment

We also deployed DKPROMPT on real robot hardware to perform object rearrangement tasks (Figure 5), where the goal is to “collect” toys using a container and place them in the middle of the table (i.e., goal area). Our real-robot setup includes a UR5e Arm with a Hand-E gripper mounted on a Segway base, and an overhead RGB-D camera (relatively fixed to the robot) for perception. We assume that the robot has a predefined set of skills, including `pick`, `place`, and `find`. `Pick` and `place` actions are implemented using GG-CNN [76], and `find` action simply uses base rotation for capturing tabletop images from different angles.

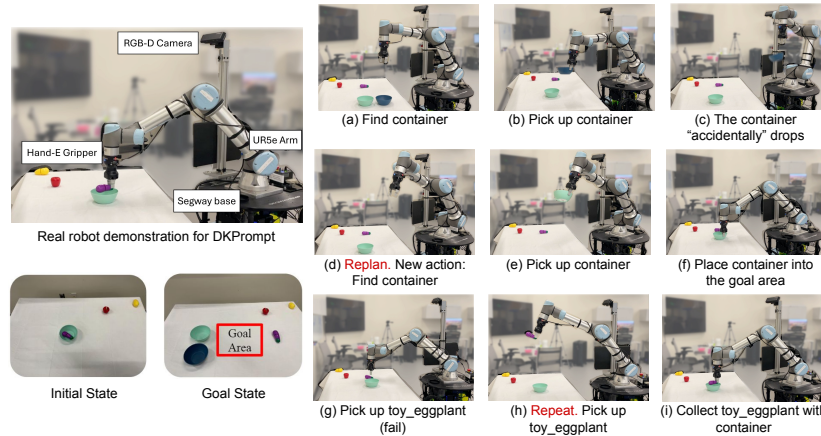


Figure 5: Screenshots showing the full demonstration trial of DKPROMPT as applied to a real robot.

246

Given the task description, the robot first decided to execute “find container” and “pick up container”. These two actions were successfully executed as shown in Figure 5(a), 5(b). When the robot was preparing for the next action (i.e., “Place container into the goal area”), the blue container accidentally dropped from the robot’s gripper to the ground (Figure 5(c)). Instead of directly executing the next action, DKPROMPT enabled the robot to check preconditions by querying the VLM “*Is the container in a robot’s hand?*”. After receiving negative feedback from the VLM, DKPROMPT updated the world state by removing `in_hand(container)` and called the planner to generate a new plan that started the task again by finding another container (Figure 5(d)). Then the robot picked up the cyan container and placed it in the middle of the table as shown in Figure 5(e), 5(f). The subsequent actions in the plan were to find and pick up a toy, but the `pick` action failed (Figure 5(g)). DKPROMPT managed to detect the failure by querying 1) “*Is there a toy\_eggplant on the table?*”, and 2) “*Is the toy\_eggplant in a robot’s hand?*”, and receiving Yes and No answers respectively. As a result, our system suggested the robot repeat the `pick` action again (Figure 5(h)). Finally, the robot successfully collected the toy by putting it into the cyan container that was previously placed in the goal area (Figure 5(i)).

## 5 Conclusion

262

In this paper, we built the synergy between classical task planning and large Vision-Language Models (VLMs), focusing on how VLMs facilitate robot planning in open-world scenarios. We propose DKPROMPT which automates VLM prompting using domain knowledge in PDDL for classical planning and task execution. Experimental results demonstrate that DKPROMPT adaptively generate visually-grounded task plans, recovers from action failures and re-plans when situations occur, outperforming classical planning, pure VLM-based and a few other competitive baselines.



## References

- [1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [2] B. Lester, R. Al-Rfou, and N. Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.
- [3] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [4] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.
- [5] M. Ghallab, D. Nau, and P. Traverso. *Automated planning and acting*. Cambridge University Press, 2016.
- [6] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, M. Lingelbach, J. Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, pages 80–93. PMLR, 2023.
- [7] Y. Ding, X. Zhang, S. Amiri, N. Cao, H. Yang, A. Kaminski, C. Esselink, and S. Zhang. Integrating action knowledge and llms for task planning and situation handling in open worlds. *Autonomous Robots*, 47(8):981–997, 2023.
- [8] N. J. Nilsson et al. Shakey the robot. 1984.
- [9] Y.-q. Jiang, S.-q. Zhang, P. Khandelwal, and P. Stone. Task planning in robotics: an empirical comparison of pddl-and asp-based systems. *Frontiers of Information Technology & Electronic Engineering*, 20:363–373, 2019.
- [10] G. Brewka, T. Eiter, and M. Truszczyński. Answer set programming at a glance. *Communications of the ACM*, 54(12):92–103, 2011.
- [11] V. Lifschitz. Answer set programming and plan generation. *Artificial Intelligence*, 138(1-2): 39–54, 2002.
- [12] M. Fox and D. Long. Pddl2. 1: An extension to pddl for expressing temporal planning domains. *Journal of artificial intelligence research*, 20:61–124, 2003.
- [13] F. Lagriffoul, N. T. Dantam, C. Garrett, A. Akbari, S. Srivastava, and L. E. Kavraki. Platform-independent benchmarks for task and motion planning. *IEEE Robotics and Automation Letters*, 3(4):3765–3772, 2018.
- [14] L. P. Kaelbling and T. Lozano-Pérez. Integrated task and motion planning in belief space. *The International Journal of Robotics Research*, 32(9-10):1194–1227, 2013.
- [15] S. Zhang, F. Yang, P. Khandelwal, and P. Stone. Mobile robot planning using action language bc with an abstraction hierarchy. In *International Conference on Logic Programming and Nonmonotonic Reasoning*, pages 502–516. Springer, 2015.
- [16] Y. Ding, X. Zhang, X. Zhan, and S. Zhang. Task-motion planning for safe and efficient urban driving. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.
- [17] Y. Jiang, H. Yedidsion, S. Zhang, G. Sharon, and P. Stone. Multi-robot planning with conflicts and synergies. *Autonomous Robots*, 43(8):2011–2032, 2019.

- [18] Y. Ding, X. Zhang, X. Zhan, and S. Zhang. Learning to ground objects for robot task and motion planning. *IEEE Robotics and Automation Letters*, 7(2):5536–5543, 2022.
- [19] Y. Zhu, J. Tremblay, S. Birchfield, and Y. Zhu. Hierarchical planning for long-horizon manipulation with geometric and symbolic scene graphs. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6541–6548. IEEE, 2021.
- [20] X. Zhang, Y. Zhu, Y. Ding, Y. Zhu, P. Stone, and S. Zhang. Visually grounded task and motion planning for mobile manipulation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 1925–1931. IEEE, 2022.
- [21] D. Driess, J.-S. Ha, and M. Toussaint. Deep visual reasoning: Learning to predict action sequences for task and motion planning from an initial scene image. *arXiv preprint arXiv:2006.05398*, 2020.
- [22] D. Driess, O. Oguz, J.-S. Ha, and M. Toussaint. Deep visual heuristics: Learning feasibility of mixed-integer programs for manipulation planning. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9563–9569. IEEE, 2020.
- [23] A. M. Wells, N. T. Dantam, A. Shrivastava, and L. E. Kavraki. Learning feasibility for task and motion planning in tabletop environments. *IEEE robotics and automation letters*, 4(2): 1255–1262, 2019.
- [24] T. Migimatsu and J. Bohg. Grounding predicates through actions. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 3498–3504. IEEE, 2022.
- [25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [26] OpenAI. Chatgpt. Accessed: 2023-02-08, 2023. URL <https://openai.com/blog/chatgpt/>. cit. on pp. 1, 16.
- [27] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- [28] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- [29] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.
- [30] Y. Kant, A. Ramachandran, S. Yenamandra, I. Gilitschenski, D. Batra, A. Szot, and H. Agrawal. Housekeep: Tidying virtual households using commonsense reasoning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*, pages 355–373. Springer, 2022.
- [31] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR, 2022.
- [32] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.
- [33] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.

- [34] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg. Progprompt: Generating situated robot task plans using large language models. *arXiv preprint arXiv:2209.11302*, 2022.
- [35] Z. Zhao, W. S. Lee, and D. Hsu. Large language models as commonsense knowledge for large-scale task planning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=WjplAYB8lH>.
- [36] W. Liu, T. Hermans, S. Chernova, and C. Paxton. Structdiffusion: Object-centric diffusion for semantic rearrangement of novel objects. In *Workshop on Language and Robotics at CoRL 2022*, 2022.
- [37] J. Wu, R. Antonova, A. Kan, M. Lepert, A. Zeng, S. Song, J. Bohg, S. Rusinkiewicz, and T. Funkhouser. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots*, 47(8):1087–1102, 2023.
- [38] K. Rana, J. Haviland, S. Garg, J. Abou-Chakra, I. Reid, and N. Suenderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable task planning. *arXiv preprint arXiv:2307.06135*, 2023.
- [39] K. Valmeekam, A. Olmo, S. Sreedharan, and S. Kambhampati. Large language models still can’t plan (a benchmark for llms on planning and reasoning about change). *arXiv preprint arXiv:2206.10498*, 2022.
- [40] K. Valmeekam, M. Marquez, S. Sreedharan, and S. Kambhampati. On the planning abilities of large language models-a critical investigation. *Advances in Neural Information Processing Systems*, 36:75993–76005, 2023.
- [41] OpenAI. Gpt-4 technical report, 2023.
- [42] T. Silver, V. Hariprasad, R. S. Shuttlesworth, N. Kumar, T. Lozano-Pérez, and L. P. Kaelbling. Pddl planning with pretrained large language models. In *NeurIPS 2022 foundation models for decision making workshop*, 2022.
- [43] V. Pallagani, B. Muppasani, K. Murugesan, F. Rossi, L. Horesh, B. Srivastava, F. Fabiano, and A. Loreggia. Plansformer: Generating symbolic plans using transformers. *arXiv preprint arXiv:2212.08681*, 2022.
- [44] D. Arora and S. Kambhampati. Learning and leveraging verifiers to improve planning capabilities of pre-trained language models. *arXiv preprint arXiv:2305.17077*, 2023.
- [45] T. Silver, S. Dan, K. Srinivas, J. B. Tenenbaum, L. Kaelbling, and M. Katz. Generalized planning in pddl domains with pretrained large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20256–20264, 2024.
- [46] Y. Chen, J. Arkin, Y. Zhang, N. Roy, and C. Fan. Autotamp: Autoregressive task and motion planning with llms as translators and checkers. *arXiv preprint arXiv:2306.06531*, 2023.
- [47] S. Wang, M. Han, Z. Jiao, Z. Zhang, Y. N. Wu, S.-C. Zhu, and H. Liu. Llm<sup>+</sup> 3: Large language model-based task and motion planning with motion failure reasoning. *arXiv preprint arXiv:2403.11552*, 2024.
- [48] B. Liu, Y. Jiang, X. Zhang, Q. Liu, S. Zhang, J. Biswas, and P. Stone. Llm+ p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*, 2023.
- [49] K. Stein and A. Koller. Autoplanbench:: Automatically generating benchmarks for llm planners from pddl. *arXiv preprint arXiv:2311.09830*, 2023.

- [50] L. Guan, K. Valmeekam, S. Sreedharan, and S. Kambhampati. Leveraging pre-trained large language models to construct and utilize world models for model-based task planning. *Advances in Neural Information Processing Systems*, 36:79081–79094, 2023.
- [51] Y. Ding, X. Zhang, C. Paxton, and S. Zhang. Task and motion planning with large language models for object rearrangement. *arXiv preprint arXiv:2303.06247*, 2023.
- [52] D. Zhang, Y. Yu, C. Li, J. Dong, D. Su, C. Chu, and D. Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024.
- [53] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [54] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [55] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [56] Anthropic. Claude 3 family, 2023. URL <https://www.anthropic.com/news/claude-3-family>. Accessed: 2024-05-21.
- [57] N. Wake, A. Kanehira, K. Sasabuchi, J. Takamatsu, and K. Ikeuchi. Gpt-4v (ision) for robotics: Multimodal task planning from human demonstration. *arXiv preprint arXiv:2311.12015*, 2023.
- [58] A. Lykov, M. Litvinov, M. Konenkov, R. Prochii, N. Burtsev, A. A. Abdulkarim, A. Bazhenov, V. Berman, and D. Tsetserukou. Cognitivedog: Large multimodal model based system to translate vision and language into action of quadruped robot. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 712–716, 2024.
- [59] L. Guan, Y. Zhou, D. Liu, Y. Zha, H. B. Amor, and S. Kambhampati. ” task success” is not enough: Investigating the use of video-language models as behavior critics for catching undesirable agent behaviors. *arXiv preprint arXiv:2402.04210*, 2024.
- [60] A. Majumdar, A. Ajay, X. Zhang, P. Putta, S. Yenamandra, M. Henaff, S. Silwal, P. Mcvay, O. Maksymets, S. Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *2nd Workshop on Mobile Manipulation and Embodied Intelligence at ICRA 2024*, 2024.
- [61] P. Sermanet, T. Ding, J. Zhao, F. Xia, D. Dwibedi, K. Gopalakrishnan, C. Chan, G. Dulac-Arnold, S. Maddineni, N. J. Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. *arXiv preprint arXiv:2311.00899*, 2023.
- [62] H. Ha and S. Song. Semantic abstraction: Open-world 3d scene understanding from 2d vision-language models. In *Conference on Robot Learning*, 2022.
- [63] L. Fan, G. Wang, Y. Jiang, A. Mandlekar, Y. Yang, H. Zhu, A. Tang, D.-A. Huang, Y. Zhu, and A. Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *arXiv preprint arXiv:2206.08853*, 2022.
- [64] N. M. M. Shafiullah, C. Paxton, L. Pinto, S. Chintala, and A. Szlam. Clip-fields: Weakly supervised semantic fields for robotic memory. *arXiv preprint arXiv: Arxiv-2210.05663*, 2022.
- [65] M. Shridhar, L. Manuelli, and D. Fox. Cliport: What and where pathways for robotic manipulation. In *Proceedings of the 5th Conference on Robot Learning (CoRL)*, 2021.

- 444 [66] A. Stone, T. Xiao, Y. Lu, K. Gopalakrishnan, K.-H. Lee, Q. Vuong, P. Wohlhart, B. Zitkovich,  
445 F. Xia, C. Finn, and K. Hausman. Open-world object manipulation using pre-trained vision-  
446 language model. In *arXiv preprint*, 2023.
- 447 [67] Q. Lv, H. Li, X. Deng, R. Shao, M. Y. Wang, and L. Nie. Robomp2: A robotic multimodal  
448 perception-planning framework with mutlimodal large language models. In *International Con-  
449 ference on Machine Learning*, 2024.
- 450 [68] X. Zhao, M. Li, C. Weber, M. B. Hafez, and S. Wermter. Chat with the environment: Inter-  
451 active multimodal perception using large language models. In *2023 IEEE/RSJ International  
452 Conference on Intelligent Robots and Systems (IROS)*, pages 3590–3596. IEEE, 2023.
- 453 [69] A. Z. Ren, A. Dixit, A. Bodrova, S. Singh, S. Tu, N. Brown, P. Xu, L. Takayama, F. Xia,  
454 J. Varley, et al. Robots that ask for help: Uncertainty alignment for large language model  
455 planners. *arXiv preprint arXiv:2307.01928*, 2023.
- 456 [70] P. Zhi, Z. Zhang, M. Han, Z. Zhang, Z. Li, Z. Jiao, B. Jia, and S. Huang. Closed-loop open-  
457 vocabulary mobile manipulation with gpt-4v, 2024.
- 458 [71] Y. Du, K. Konyushkova, M. Denil, A. Raju, J. Landon, F. Hill, N. de Freitas, and S. Cabi.  
459 Vision-language models as success detectors. *arXiv preprint arXiv:2303.07280*, 2023.
- 460 [72] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson,  
461 Q. Vuong, T. Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint  
462 arXiv:2303.03378*, 2023.
- 463 [73] Y. Guo, Y.-J. Wang, L. Zha, Z. Jiang, and J. Chen. Doremi: Grounding language model by  
464 detecting and recovering from plan-execution misalignment. *arXiv preprint arXiv:2307.00329*,  
465 2023.
- 466 [74] M. Helmert. The fast downward planning system. *Journal of Artificial Intelligence Research*,  
467 26:191–246, 2006.
- 468 [75] M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. Lillicrap, J.-b. Alayrac, R. Soricut,  
469 A. Lazaridou, O. Firat, J. Schrittwieser, et al. Gemini 1.5: Unlocking multimodal under-  
470 standing across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- 471 [76] D. Morrison, P. Corke, and J. Leitner. Closing the loop for robotic grasping: A real-time,  
472 generative grasp synthesis approach. *arXiv preprint arXiv:1804.05172*, 2018.




473 **Appendix**

474 **[todo: I think we should highlight our webpage again in the appendix since we plan to include our**  
 475 **video in the website instead of submitting it with supp.]**

DKPROMPT

**System:** Imagine you are an intelligent agent that can answer questions based on what you see. You will be given a single image as the agent’s current view, and one or more yes/no question(s) asking about the image. Questions will be separated by semicolon. For each question, you should answer "yes", "no", or "skip" without any explanation. Answer "yes" or "no" only if you are pretty sure about what you see in the image. It’s fine to answer "skip" to skip the question if you are not confident about your answer. Answers should be separated by semicolon (e.g., "yes;no;skip" for three questions).

**DKPROMPT:** Is cup inview agent?;Is cup inside cabinet?;Is cabinet open?



476

Table 5: Full results with the total numbers of trials and successful trials.

#	Methods	Tasks					avg. (%)	
		boil water in the microwave	bring in empty bottle	cook a frozen pie	halve an egg	store firewood		
<b>Ours</b>								
1	DKPROMPT (w/ GPT-4)	12/18	18/20	14/23	15/20	8/20	<b>66.5</b>	
<b>Baselines in Literature</b>								
2	VLM-planner	1/20	2/34	0/20	0/19	0/22	2.2	
3	Classical-planner	5/35	3/11	4/30	8/30	1/25	17.1	
4	Aff.-QA	4/28	11/20	7/20	10/20	5/20	35.9	
5	Suc.-QA	7/18	18/20	10/20	12/20	4/20	51.8	
6	Suc.Aff.-QA	8/20	12/16	11/20	13/20	7/20	54.0	
<b>Ablation</b>								
7	Eff.-only	15/30	15/16	4/15	10/15	7/25	53.0	
8	Pre.-only	3/17	15/20	7/20	11/20	5/20	41.5	
<b>Other VLMs</b>								
9	DKPROMPT (w/ Gemini-1.5)	7/20	14/20	6/20	9/20	8/20	44.0	
10	DKPROMPT (w/ Claude-3)	5/20	12/28	4/14	10/20	3/13	33.9	

477 model ckpt: claude: claude-3-opus-20240229 gemini: gemini-1.5-pro gpt-4: gpt-4-turbo

```

(define (domain omnigibson)

  (:requirements :strips :typing :negative-preconditions :conditional-
    effects)

  (:types
    movable liquid furniture room agent - object

    wooden_stick tupperware brownie beer_bottle water_bottle mug pie
      carving_knife hard__boiled_egg - movable
    water - liquid
    countertop electric_refrigerator oven cabinet sink floor
      microwave table - furniture

    kitchen living_room - room

    water-n-06 - water
    mug-n-04 - mug
    cabinet-n-01 - cabinet
    sink-n-01 - sink
    floor-n-01 - floor
    microwave-n-02 - microwave
    pie-n-01 - pie
    oven-n-01 - oven
    electric_refrigerator-n-01 - electric_refrigerator
    carving_knife-n-01 - carving_knife
    countertop-n-01 - countertop
    hard__boiled_egg-n-01 - hard__boiled_egg
    water_bottle-n-01 - water_bottle
    beer_bottle-n-01 - beer_bottle
    brownie-n-03 - brownie
    tupperware-n-01 - tupperware
    wooden_stick-n-01 - wooden_stick
    table-n-02 - table

    agent-n-01 - agent
  )

  (:predicates
    (inside ?o1 - object ?o2 - object)
    (insource ?s - sink ?w - liquid)
    (inroom ?o - object ?r - room)
    (inhand ?a - agent ?o - object)
    (inview ?a - agent ?o - object)
    (handempty ?a - agent)
    (closed ?o - object)
    (filled ?o - movable ?w - liquid)
    (filledsink ?s - sink ?w - liquid)
    (turnedon ?o - object)
    (cooked ?o - object)
    (found ?a - agent ?o - object)
    (frozen ?o - object)
    (hot ?o - object)
    (halved ?o - object)
    (onfloor ?o - object ?f - floor)
    (ontop ?o1 - object ?o2 - object)
  )

  (:action find
    :parameters (?a - agent ?o - object ?r - room)
    :precondition (and (inroom ?a ?r) (inroom ?o ?r))
    :effect (and (inview ?a ?o) (found ?a ?o) (forall
      (?oo - object)
      (when
        (found ?a ?oo)
        (not (found ?a ?oo))))))
  )

```

Figure 6

```

(:action graspon
  :parameters (?a - agent ?o1 - movable ?o2 - object)
  :precondition (and (inview ?a ?o1) (found ?a ?o1) (handempty ?a)
    (ontop ?o1 ?o2))
  :effect (and (not (inview ?a ?o1)) (not (handempty ?a)) (inhand ?
    a ?o1) (not (ontop ?o1 ?o2)))
)

(:action graspin
  :parameters (?a - agent ?o1 - movable ?o2 - object)
  :precondition (and (inview ?a ?o1) (found ?a ?o1) (handempty ?a)
    (inside ?o1 ?o2))
  :effect (and (not (inview ?a ?o1)) (not (handempty ?a)) (inhand ?
    a ?o1) (not (inside ?o1 ?o2)))
)

(:action placein
  :parameters (?a - agent ?o1 - movable ?o2 - object)
  :precondition (and (not (handempty ?a)) (inhand ?a ?o1) (inview ?
    a ?o2) (found ?a ?o2) (not (closed ?o2)))
  :effect (and (handempty ?a) (not (inhand ?a ?o1)) (inside ?o1 ?o2
    ) (forall
      (?oo - object)
      (when
        (inside ?oo ?o1)
        (inside ?oo ?o2))
    ))
)

(:action placeon
  :parameters (?a - agent ?o1 - movable ?o2 - object)
  :precondition (and (not (handempty ?a)) (inhand ?a ?o1) (inview ?
    a ?o2) (found ?a ?o2))
  :effect (and (handempty ?a) (not (inhand ?a ?o1)) (ontop ?o1 ?o2)
    )
)

(:action fillsink
  :parameters (?a - agent ?s - sink ?w - liquid)
  :precondition (and (inview ?a ?s) (found ?a ?s) (insource ?s ?w))
  :effect (filledsink ?s ?w)
)

(:action fill
  :parameters (?a - agent ?o - movable ?s - sink ?w - liquid)
  :precondition (and (inhand ?a ?o) (not (handempty ?a)) (
    filledsink ?s ?w) (inview ?a ?s) (found ?a ?s))
  :effect (and (filled ?o ?w) (not (filledsink ?s ?w)))
)

(:action openit
  :parameters (?a - agent ?o - object ?r - room)
  :precondition (and (inview ?a ?o) (found ?a ?o) (inroom ?o ?r))
  :effect (and (not (closed ?o)) (forall
    (?oo - object)
    (when
      (inside ?oo ?o)
      (inroom ?oo ?r))
    ))
)

```

Figure 7

```

(:action closeit
:parameters (?a - agent ?o - object ?r - room)
:precondition (and (inview ?a ?o) (found ?a ?o) (inroom ?o ?r))
:effect (and (closed ?o) (forall
  (?oo - object)
  (when
    (inside ?oo ?o)
    (not (inroom ?oo ?r)))
  ))
)

(:action microwave_water
:parameters (?a - agent ?m - microwave ?o - movable ?w - water)
:precondition (and (inview ?a ?m) (found ?a ?m) (closed ?m) (
  inside ?o ?m) (filled ?o ?w))
:effect (and (turnedon ?m) (cooked ?w))
)

(:action heat_food_with_oven
:parameters (?a - agent ?v - oven ?f - object)
:precondition (and (inview ?a ?v) (found ?a ?v) (inside ?f ?v))
:effect (and (hot ?f) (turnedon ?v))
)

(:action cut_into_half
:parameters (?a - agent ?k - carving_knife ?o - object)
:precondition (and (inview ?a ?o) (found ?a ?o) (not (handempty ?
  a)) (inhand ?a ?k))
:effect (halved ?o)
)

(:action place_on_floor
:parameters (?a - agent ?o - object ?f - floor)
:precondition (and (inview ?a ?f) (found ?a ?f) (not (handempty ?
  a)) (inhand ?a ?o))
:effect (and (handempty ?a) (not (inhand ?a ?o)) (onfloor ?o ?f))
)
)

```

Figure 8

Table 6: Actions and their situation parameters.

Actions	Uncertain outcomes	
	Situations	Probabilities
find	(1) The robot succeeds in navigation but the object is not inview.	N/A
	(2) There is no free space near the object so navigation fails.	N/A
	(3) The object that the robot is holding drops during navigation.	0.1
grasp	(1) The robot fails to grasp, and the object position remains unchanged.	0.25
	(2) The robot fails to grasp, and the object drops nearby.	0.25
placein	(1) The robot fails to place, and the object remains in the robot's hand.	0.1
	(2) The robot fails to place, and the object drops nearby.	0.1
placeon	(1) The robot fails to place, and the object remains in the robot's hand.	0.1
	(2) The robot fails to place, and the object drops nearby.	0.1
fillsink	(1) The robot fails to open the faucet.	0.1
fill	(1) The container is not fully filled.	0.05
	(2) The container drops nearby.	0.05
open	(1) The robot fails to open and the object remains closed.	0.1
close	(1) The robot fails to close and the object remains open.	0.1
turnon	(1) The robot fails to turn on the switch and the object remains off.	0.1
cut	(1) The object is not cut into half, and the knife is still in the robot's hand.	0.25
	(2) The object is not cut into half, and the knife drops nearby.	0.25

Table 7: Model checkpoints we used for off-the-shelf VLMs.

<b>VLM</b>	<b>Model</b>
GPT-4	gpt-4-turbo
Gemini	gemini-1.5-pro
Claude	claude-3-opus-20240229